PB81-112534

Validation of Air Monitoring Data

PEDCo Environmental, Inc.
Durham, NC

Prepared for

Environmental Monitoring Systems Lab.
Research Triangle Park, NC

Jun 80

U.S. Department of Commerce
National Technical Information Service
NTIS

United States
Environmental Protection
Agency

Environmental Monitoring and Support
Laboratory
Research Triangle Park NC 27711

EPA-600/4-80-030
June 1980

Research and Development

PB81-112534

EPA

Validation of Air Monitoring Data

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/4-80-030	2.
3. RECIPIENT'S ACCESSION NO. PB81 112534	
4. TITLE AND SUBTITLE: VALIDATION OF AIR MONITORING DATA	5. REPORT DATE: June 1980
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): A. Carl Nelson, Jr., Dave W. Armentrout, and Ted R. Johnson	8. PERFORMING ORGANIZATION REPORT NO. 3320-N
9. PERFORMING ORGANIZATION NAME AND ADDRESS: PEDCo Environmental, Inc., 505 South Duke Street, Suite 503, Durham, North Carolina 27701	10. PROGRAM ELEMENT NO. Assignment No. 14
	11. CONTRACT/GRANT NO. 68-02-2722
12. SPONSORING AGENCY NAME AND ADDRESS: Office of Research and Development, Environmental Monitoring Systems Laboratory, Research Triangle Park, N. C. 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA 600/08

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Data validation refers to those activities performed after the data have been obtained and thus serves as a final screening of the data before they are used in a decision making process. This report provides organizations that are monitoring ambient air levels and stationary source emissions with a collection of data validation procedures and with criteria for selection of the appropriate procedures for the particular application. Both hypothetical and case studies, and several examples are given to illustrate the use of the procedures. Statistical procedures and tables are in the appendices.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Data Validation Data Screening Data Editing Quality Assurance Outliers Statistics Environmental Data	Environmental Monitoring Data Management —	43F 68A

18. DISTRIBUTION STATEMENT: Release to public	19. SECURITY CLASS (This Report): Unclassified	21. NO. OF PAGES
	20. SECURITY CLASS (This page): Unclassified	22. PRICE

EPA Form 2220-1 (9-73)

EPA-600/4-80-030

June 1980

Validation of Air Monitoring Data

by

A. Carl Nelson, Jr., Dave W. Armentrout
and Ted R. Johnson

PEDCo Environmental, Inc.
Durham, North Carolina 27701

Contract No. 68-02-2722
Task No. 14

EPA Project Officer: Raymond C. Rhodes
Data Management and Analysis Division

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Environmental Monitoring Systems Laboratory
Research Triangle Park, North Carolina 27711

DISCLAIMER

This report has been reviewed by the Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ACKNOWLEDGEMENT

This report was prepared for the Environmental Monitoring Systems Laboratory of the U.S. Environmental Protection Agency. Mr. Raymond C. Rhodes was the Project Officer. PEDCo Environmental, Inc., appreciates the direction and extensive review provided. Appreciation is also expressed to Ms. Debora Pizer and Mr. José Suñé for the information on CHAMP, to Mr. Robert B. Jurgens for information on RAMS, to Mr. E. Gardner Evans for examples of data validation tests and to Mr. Seymour Hochheiser for discussions and reviews. In the preparation of this report, materials from the Data Validation lectures of the Air Pollution Training Institute (APTI), 470 course (Quality Assurance for Air Pollution Measurement Systems) and examples from the Office of Air Quality Planning and Standards (OAQPS) guideline have been used. Mr. A. Carl Nelson served as PEDCo's Project Manager. Mr. Nelson, Mr. Dave W. Armentrout, and Mr. Ted R. Johnson were the principal authors.

FOREWORD

Measurement and monitoring research efforts are designed to anticipate potential environmental problems, to support regulatory actions by developing an in-depth understanding of the nature and processes that impact health and the ecology, to provide innovative means of monitoring compliance with regulations and to evaluate the effectiveness of health and environmental protection efforts through the monitoring of long-term trends. The Environmental Monitoring Systems Laboratory, Research Triangle Park, North Carolina, has the responsibility for: assessment of environmental monitoring technology and systems; implementation of agency-wide quality assurance programs for air pollution measurement systems; and supplying technical support to other groups in the Agency including the Office of Air, Noise and Radiation, the Office of Toxic Substances, and the Office of Enforcement.

Data validation, an element of quality assurance, is necessary to provide accurate and reliable environmental data. Data of known and acceptable quality are needed for measuring compliance with regulations, assessing health effects, and developing optimum strategies to cope with environmental pollution situations. A unified treatment of validation of particular types of data bases is needed to support broad-scale uses of these data. This report presents a systematic review of procedures and techniques which have proven useful in performing the data validation function. Recommendations are given for the selection and implementation of procedures and techniques appropriate for organizations performing air monitoring and depending upon their particular mode of operations, computational and statistical capability, and monitoring objectives.

ABSTRACT

Data validation refers to those activities performed after the data have been obtained and thus serves as a final screening of the data before they are used in a decision making process. This report provides organizations that are monitoring ambient air levels and stationary source emissions with a collection of data validation procedures and with criteria for selection of the appropriate procedures for the particular application. Both hypothetical and case studies, and several examples are given to illustrate the use of the procedures. Statistical procedures and tables are in the appendices.

CONTENTS

	Page
Acknowledgment	iii
Abstract	v
Figures	ix
Tables	x
Executive Summary	
1.0 Introduction	1-1
1.1 Purpose	1-1
1.2 Scope and Organization of the Document	1-1
1.3 References	1-2
2.0 Background	2-1
2.1 Definition and Scope of Data Validation	2-1
2.2 Data Validation Procedures	2-2
2.3 Selection of Data Validation Procedures	2-2
2.4 Implementation of Data Validation	2-3
2.5 Brief Literature Review	2-3
2.6 References	2-4
3.0 Data Validation Procedures	3-1
3.1 Routine Procedures	3-1
3.1.1 Data identification checks	3-2
3.1.2 Unusual event review	3-2
3.1.3 Deterministic relationship checks	3-2
3.1.4 Data processing procedures	3-3
3.2 Tests for Internal Consistency	3-5
3.2.1 Data plots	3-5
3.2.2 Dixon ratio test	3-6
3.2.3 Grubbs test	3-12
3.2.4 Gap test	3-14
3.2.5 "Johnson" p test	3-16
3.2.6 Multivariate tests	3-17
3.3 Tests for Historical Consistency	3-19
3.3.1 Gross limit checks	3-19
3.3.2 Pattern and successive value tests	3-21
3.3.3 Parameter relationship test	3-21
3.3.4 Shewhart control chart	3-23

CONTENTS (continued)

		Page
3.4	Tests for Consistency of Parallel Data Sets	3-29
3.4.1	The sign test	3-29
3.4.2	Wilcoxon signed-rank test	3-32
3.4.3	Rank sum test	3-34
3.4.4	Intersite correlation test	3-38
3.5	References	3-44
4.0	Selection and Implementation of Procedures	4-1
4.1	Organizational Criteria	4-2
4.1.1	Number of data sets	4-2
4.1.2	Historical data requirements	4-3
4.1.3	Nature of data anomalies	4-3
4.1.4	Strip chart data	4-3
4.1.5	Size of data set	4-4
4.1.6	Magnetic tape data	4-4
4.1.7	Data transmitted by telephone lines	4-4
4.1.8	Timeliness of procedure	4-5
4.2	Analytical Criteria	4-5
4.2.1	Statistical sophistication	4-5
4.2.2	Computational requirements	4-5
4.2.3	Expense of analysis	4-6
4.2.4	Sensitivity of the procedure	4-6
4.2.5	Use of data	4-7
4.3	Implementation of Data Validation Process	4-7
4.4	References	4-9
5.0	Hypothetical Examples and Case Studies	5-1
5.1	Hypothetical Example for Ambient Air Monitoring	5-1
5.1.1	Without computerized support	5-1
5.1.2	With computerized support	5-4
5.2	Hypothetical Example for Source Testing	5-4
5.3	Case Study of a Manual Data Validation System	5-5
5.3.1	Quality control	5-6
5.3.2	Data validation	5-8

CONTENTS (continued)

	Page
5.4 Case Study of the CHAMP Automated Data Validation System	5-10
5.4.1 Quality control functions	5-12
5.4.2 Data validation	5-14
5.5 Case Study of a Regional Air Monitoring System (RAMS) Data Validation	5-23
5.5.1 Quality control	5-23
5.5.2 Data validation	5-27
5.6 References	5-35
6.0 Bibliography	6-1
Appendix A - Statistical Tables	A-1
Appendix B - Fitting Distributions to Data	B-1
Appendix C - Calculation of Limits for Shewhart Control Chart	C-1

FIGURES

Number		Page
3-1	Computer plot of data with single anomalous value	3-7
3-2	Computer plot of data with single anomalous value	3-8
3-3	Illustration of Dixon ratio test for two TSP monthly data sets	3-10
3-4	Illustration of Grubbs test for two TSP monthly data sets	3-13
3-5	Illustration of gap test for hourly ozone data (Newport, KY, 1976)	3-15
3-6	P(x) versus concentration for Weibull distribution fitted to 1976 NO_2 data from Essex, MD. Acceptance range defined as $0.05 < P(x) < 0.95$	3-18
3-7	Shewhart control chart for mean values with 1978 data plotted	3-26
3-8	Shewhart control chart for range values with 1978 data plotted	3-27
3-9	Intersite correlation test data	3-42
5-1	Automated quality control tests	5-13
5-2	Example CHAMP data validation report (partial printout)	5-16
5-3	High/low critical values for CHAMP secondary parameters (partial printout)	5-16
5-4	CHAMP validation system, invalidity causes by hour	5-17
5-5	Five-minute values of invalid secondary parameters	5-17
5-6	Example CHAMP journal entries for data validation	5-18
5-7	Example CHAMP validation data review	5-19
5-8	Example ozone daily data curve from CHAMP data validation	5-21
5-9	Example NO_x curve from CHAMP data validation	5-22
5-10	Location of RAMS stations	5-25
5-11	RAMS data flow: RAMS station	5-26
5-12	RAMS data flow, central facility	5-28
5-13	RAMS hourly average temperature data, January 21, to February 9, 1975	5-31
5-14	Minute temperature data from Station 103, from 1200 to 2359 hours, February 1, 1975	5-32
5-15	Reference weather data used in RAMS data validation	5-34
5-16	Examples of instrument responses that can be detected through minute successive differences	5-35

TABLES

Number		Page
3-1	Examples of Hourly and Daily Gross Limit Checks for Ambient Air Monitoring	3-20
3-2	Summary of Limit Values used in EPA Region V for Pattern Tests	3-22
3-3	TSP Data from Site 397140014H01 Selected as Historical Data Base for Shewhart Control Chart	3-25
3-4	TSP Data from Site 397140014H01 for Control Chart (1978)	3-25
3-5	Ozone Data (ppb) Recorded August 1, 1978, at Monitoring Sites 909929911I01 (Site A) and 090020013I01 (Site B)	3-31
3-6	Ozone Data (ppb) Incorporating Simulated +5 ppb Calibration Error at Site A	3-33
3-7	Procedure for Wilcoxon Signed-Rank Test	3-34
3-8	Application of Wilcoxon Signed-Rank Test to Ozone Data (ppb) in Table 3-4	3-35
3-9	Procedure for Wilcoxon Rank Sum Test	3-36
3-10	Application of Rank Sum Test to Ozone Data in Table 3-5	3-37
3-11	TSP Data from Sites 397140014H01 and 397140020H01 for Intersite Correlation Test, 1978	3-40
3-12	χ^2 Values for Two Degrees of Freedom, for Various Probability Levels	3-43
4-1	Factors to Consider in the Selection of Data Validation Procedures	4-1
4-2	Selection of Data Validation Procedures	4-2
5-1	CHAMP Environmental Parameters	5-11
5-2	Parameters Monitored in the Regional Air Monitoring System	5-24
5-3	Typical Gross Limit Values used in the Regional Air Monitoring System Data Validation Procedure	5-29
A-1	Dixon Criteria for Testing of Extreme Observation (Single Sample)	A-1
A-2	Critical Values for 5% and 1% Tests of Discordancy for Two Outliers in a Normal Sample	A-2
A-3	Critical T Values for One-Sided Grubbs Test when Standard Deviation is Calculated from Sample	A-3
A-4	Wilcoxon Signed-Rank Test	A-6
A-5	Rank Sum Test α = P[H_0 is true]	A-7
B-1	Estimation of Distribution Parameters	B-4
C-1	Factors for Estimating Control Limits of Shewhart Chart	C-2

EXECUTIVE SUMMARY

An essential element of quality assurance is the validation of data. The data validation procedures in this report span the range of application from the collectors of data (e.g., State or local agencies) to the users of the data (researchers of regional or Federal agencies, consultants, and industrial personnel).

It is stressed that data validation is most effectively performed where the data are collected; questionable data can be most easily checked at this level. The validation procedures are described in increasing order of complexity and statistical sophistication. They are grouped into four categories for ease in understanding their role in applications: procedures which should be applied routinely, those which are used to test the internal consistency within a given data set, and procedures for comparing data sets with historical data and with other data sets. The procedures applicable at the originating level are described first. The more complex statistical procedures may be used if the level of computer capability, training, and support is adequate.

The proper implementation of data validation will ensure its effectiveness. An individual should be assigned the responsibility of data validation, even though it may be a part-time activity. One responsibility will be to develop the data validation plan as an integral part of the agency quality assurance plan. A data flow diagram should be developed to identify each step of data handling process which may result in an error or in lost data. The data validation plan can be developed from the techniques available in this report or in the literature (many references are included in this report). A review and evaluation of the effectiveness of the data validation process should be made periodically 1) to optimize the sensitivity of the checks, and 2)

to summarize and evaluate the basic reasons and causes of invalidation. Validation limits may be changed to alter the sensitivity. A particular procedure may not be adequate or may be too costly (relative to its effectiveness) to be used on a routine basis.

In summary, data validation is an effective means of ensuring the integrity of the data for use in decisionmaking.

1.0 INTRODUCTION

1.1 PURPOSE

The purpose of this report is to provide organizations that are monitoring ambient air levels and stationary source emissions with (1) a collection of useful data validation procedures and with (2) criteria for selecting the appropriate procedures for the application. Data validation procedures are an integral part of a complete quality assurance program.

In this report, data validation will refer to those activities performed after the fact--that is, after the data have been obtained. Quality control has the purpose of minimizing the amount of bad data collected or obtained. Data validation is to prevent the remaining bad data from getting through the data collection and storage system. Data validation thus serves as a final screen before the data are used in the decision making process. Whether data validation is performed by a data validator with this specific assignment, by a researcher using an existing data bank, or by a member of a field team or local agency, it is preferable that data validation be performed as soon as possible after the data have been obtained. At this stage, the questionable data can be checked by recalling information concerning unusual events and meteorological conditions which can aid in the validation process. Also, timely corrective actions may be taken when indicated to minimize further generation of questionable data.

1.2 SCOPE AND ORGANIZATION OF THE DOCUMENT

Although this report includes discussion of the general theory of data validation, the examples described here concern air pollution data. In particular, the validation procedures can be applied to ambient air monitoring data, source test data, and meteorological data. This document presents general data validation procedures in Section 3, guidelines for selection and

implementation of data validation procedures in Section 4, and case studies and hypothetical examples in Section 5. The references are given at the end of each section. Appendices A, B, and C contain tables and supplemental mathematical and statistical background. The reader will not need a mathematical or statistical background to follow the guidelines in this document; he or she needs only an interest in developing a data validation process.

This document is not intended to replace previous publications. Some recent publications are recommended to the reader for additional information.[1-5] Section 2 contains additional background information.

1.3 REFERENCES

1. U.S. Environmental Protection Agency. Screening Procedures for Ambient Air Quality Data. EPA-450/2-78-037, July 1978.

2. Rhodes, R. C., and S. Hochheiser. Data Validation Conference Proceedings. Presented by Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, EPA 600/9-79-042, September 1979.

3. U.S. Department of Commerce. Computer Science and Technology: Performance Assurance and Data Integrity Practices. National Bureau of Standards, Washington, D.C., January 1978.

4. Barnett, V., and T. Lewis. Outliers in Statistical Data. John Wiley and Sons, New York, 1978.

5. Naus, J. I. Data Quality Control and Editing. Marcel Dekker, Inc., New York, 1975.

2.0 BACKGROUND

2.1 DEFINITION AND SCOPE OF DATA VALIDATION

Data validation is "a systematic process for reviewing a body of data against a set of criteria to provide assurance that the data are adequate for their intended use."[1] In technical literature, data validation may be referred to as data editing, data screening, data checking, data auditing, data verification, data certification, and technical data review. Each of these terms refers to a set of procedures for detecting, evaluating, and correcting erroneous or questionable data. Data validation is one of the primary activities for ensuring the reporting and use of good data, whereas quality control (QC) activities are designed for the purpose of acquiring good data.

The procedures included in data validation vary with the method of obtaining data. In one application the data may be recorded on a field data form and returned to a laboratory for further handling prior to storage in a data bank. In another application, such as CHAMP (Section 5), the data are obtained directly from the monitor; several data validation checks (including plotting the data) are made using the computer system, and questionable data are identified along with information on ancillary data needed by the validator to aid in the data validation.

It becomes more difficult to distinguish between QC and data validation activities when using a computerized system for acquiring, processing, checking, and finally storing the data for retrieval. However, all procedures applied after the data are first obtained and which result in the identification of questionable data and which result in subsequent investigation of these data, will be considered as data validation. The QC process will apply to the techniques applied prior to obtaining the data, such as calibration checks of the instruments and the use of control samples. Although these QC techniques may result

in data being questioned and eliminated, these data are screened by the QC system of checks and not by the data validation system. Recommended QC checks and guidelines for their use are contained in the Quality Assurance Handbook. This handbook discusses data validation briefly in Volume I and in the introduction to Volume II.[1]

The data validation process includes not only the identification or flagging of questionable data but also the investigation of apparent anomalies. The latter step is often performed by a person other than the one performing the first step, particularly when the data validation is being performed at an organizational level removed from the source of the data.

2.2 DATA VALIDATION PROCEDURES

Specific data validation procedures are described in Section 3. Although the examples and terminology used are primarily from air pollution, these procedures are not limited to this one data category. A given procedure may be used for air monitoring data (e.g., hi-vol data), for meteorological data, for source test data, and in fact for almost any data obtained by a similar measurement system.

2.3 SELECTION OF DATA VALIDATION PROCEDURES

In addition to general descriptions of the data validation procedures, this document discusses how to select the most appropriate procedure for a particular application. The procedure(s) most appropriate for a local agency with five monitoring stations and one station reporting meteorological data may not be appropriate for the performance of a Reference Method 6 source test at a utility plant. A State or Federal agency with a data bank storage and retrieval system may require still other procedures. In each application, the data validation procedures should be selected with respect to considerations such as the volume of data, type of data output/transmission, computational capability, graphics capability, and the nature of expected errors. These factors are considered in Section 4.

2.4 IMPLEMENTATION OF DATA VALIDATION

The selection of the data validation procedures is followed by the implementation of the data validation process. Decisions must be made concerning:

1. Personnel performing the procedures,
2. Frequency of data checks,
3. Methods of flagging data,
4. Procedures for investigating anomalies, and
5. Summarization of data validation including a summary of and subsequent decisions concerning flagged data.

Criteria for making these decisions are in Section 4.3.

2.5 BRIEF LITERATURE REVIEW

There are several good references on the application of validation procedures to specific types of data.[2-12] Reference 2 contains a discussion of the Dixon ratio test, the Shewhart quality control test, pattern tests for four pollutants, copies of three technical papers, a comparison of these tests, and computer programs for three test procedures (gap test, pattern test and Shewhart control chart). This development work done in EPA, Office of Air Quality Planning and Standards (OAQPS) became the basis for the Air Data Screening System, which is now being implemented in 27 states through the Air Quality Data Handling System (AQDHS). References 3 through 7 contain background information on the statistical test procedures in the OAQPS Guideline.[2] These references demonstrate and compare the effectiveness of the screening procedures of reference 2. In particular, Reference 3 describes the application and evaluation of the Dixon ratio test. Reference 4 compares seven screening test procedures using continuous 1-hour measurements for three pollutants and two tests (Dixon ratio and Shewhart control chart) using 24-hour data for three pollutants. Reference 5 compares the Dixon ratio test and the Shewhart control chart; good agreement was indicated and use of both procedures was recommended; however, the control chart procedure was preferred if only one procedure was used. Reference 6 describes the application of the control chart proce-

dure to data from monitoring sites in Region V. Reference 7 describes an automated pattern test procedure and includes a computer program for same. The limit values used in this pattern test are in Section 3.3. Reference 8 contains a collection of papers on a variety of applications in air pollution. Reference 9 is a recent publication on data validation as applied to computer systems. A major portion of the literature on data validation would be classified as statistical in content. Furthermore, many of the procedures for identifying possible data anomalies are included throughout the statistical literature under the subject of outliers. Reference 10, a recent text on this subject, summarizes the results dispersed throughout the literature. A monograph on statistical applications to data QC and editing is Reference 11. Reference 12 contains several statistical tests for outliers and the corresponding tables. Several additional references will be given at the ends of respective sections.

2.6 REFERENCES

1. U.S. Environmental Protection Agency. Quality Assurance Handbook: Vol. I, Principles; Vol. II, Ambient Air Specific Methods; and Vol. III, Stationary Source Specific Methods. EPA-600/9-76-005, 1976.

2. U.S. Environmental Protection Agency. Screening Procedures for Ambient Air Quality Data. EPA-450/2-78-037, July 1978.

3. W. F. Hunt, Jr., T. C. Curran, N. H. Frank, and R. B. Faoro, "Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air," Transactions of the Joint European Organization for Quality Control/ International Academy for Quality Conference, Vernice Lido, Italy, September 1975.

4. W. F. Hunt, Jr., R. B. Faoro, T. C. Curran, and W. M. Cox, "The Application of Quality Control Procedures to the Ambient Air Pollution Problem in the USA," Transactions of the European Organization for Quality Control, Copenhagen, Denmark, June 1976.

5. W. F. Hunt, Jr., R. B. Faoro, and S. K. Goranson, "A Comparison of the Dixon Ratio Test and Shewhart Control Test Applied to the National Aerometric Data Bank," Transactions of the American Society for Quality Control, Toronto, Canada, June 1976.

6. W. F. Hunt, Jr., J. B. Clark, and S. K. Goranson, "The Shewhart Control Chart Test: A Recommended Procedure for Screening 24-Hour Air Pollution Measurements," J. Air Poll. Control Assoc. 28:508 (1979).

7. R. B. Faoro, T. C. Curran, and W. F. Hunt, Jr., "Automated Screening of Hourly Air Quality Data," Transactions of the American Society for Quality Control, Chicago, Ill., May 1978.

8. Rhodes, R. C., and S. Hochheiser. Data Validation Conference Proceedings. Presented by Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, EPA 600/9-79-042, September 1979.

9. U.S. Department of Commerce. Computer Science and Technology: Performance Assurance and Data Integrity Practices. National Bureau of Standards, Washington, D.C., January 1978.

10. Barnett, V. and T. Lewis. Outliers in Statistical Data. John Wiley and Sons, New York, 1978.

11. Naus, J. I. Data Quality Control and Editing. Marcel Dekker, Inc., New York, 1975.

12. 1978 Annual Book of ASTM Standards, Part 41. Standard Recommended Practice for Dealing with Outlying Observations, ASTM Designation: E 178-75. pp. 212-240.

3.0 DATA VALIDATION PROCEDURES

This section contains descriptions, along with examples, of recommended data validation procedures. These procedures fall into four categories:

1. Check and review procedures which should be used to some extent in every validation process,
2. Procedures for testing the internal consistency of a single data set,
3. Procedures for testing the consistency of data sets with previous data (historical or temporal consistency), and
4. Procedures for testing the consistency of two or more data sets collected at the same time or under similar conditions (consistency of parallel data sets).

These four categories are described in Sections 3.1 to 3.4. In each section, the procedures are arranged in increasing order of statistical complexity. Hence the user of this report wishing to use the simplest procedures should refer to the first one or two procedures in each subsection. In particular, a local or State agency, with a small staff but without computer facilities and statistical support, would probably use only the procedures described in Sections 3.1.1, 3.1.2, 3.1.3, 3.2.1, 3.3.1, 3.3.2, 3.3.3, and 3.3.4. The selection of appropriate data validation procedures for a particular application is in Section 4. Implementation of the data validation process is described in Section 4.3.

3.1 ROUTINE PROCEDURES

Validation checks which should be made routinely during the processing of data include checks for proper data identification codes, review of unusual events, deterministic relationship checks, and performance checks of the data processing system.

3.1.1 Data Identification Checks

Data with improper identification codes are useless. Identification fields which must be correct are: 1) time (start and stop time and date), 2) location, 3) sampling/analytical method code, 4) pollutant method interval unit, 5) parameter, and 6) decimal. Examples of data identification problems that have been noted by the EPA regional offices include: 1) improper State identification codes; 2) data identified for a nonexistent day (e.g., October 35); and 3) duplicate data from one monitoring site but no data from another. Since most of these types of problems are the result of human error, an individual other than the original person preparing the forms should scan the data coding forms prior to using the data for computer entry or manual summary. The data listings should also be checked after entry into a computer system or data bank.

3.1.2 Unusual Event Review

A log should be maintained by each agency to record extrinsic events (e.g., construction activity, duststorms, unusual traffic volume, and traffic jams) that could explain unusual data. Depending on the purpose of data collection, this information could also be used to explain why no data are reported for a specified time interval, or it could be the basis for deleting data from a file for specific analytical purposes.

3.1.3 Deterministic Relationship Checks

Data sets which contain two or more physically or chemically related parameters should be routinely checked to ensure that the measured values on an individual parameter do not exceed the measured values of an aggregate parameter which includes the individual parameter. For example, NO_2 values should not exceed NO_x values recorded at the same time and location. The following table lists some, but not all, of the possible deterministic relationship checks involving air quality and meteorological parameters. The measured values of the individual parameters (first column of table) should not exceed the corresponding measured values of the aggregate parameter (second column).

Individual parameter	Aggregate parameter
NO (nitric oxide)	NO_x (total nitrogen oxides)
NO_2 (nitrogen dioxide)	NO_x (total nitrogen oxides)
CH_4 (methane)	THC (total hydrocarbons)
SO_2 (sulfur dioxide)	total sulfur
H_2S (hydrogen sulfide)	total sulfur
Pb (lead)	TSP (total suspended particulate)
dewpoint	temperature

Data sets in which individual parameter values exceed the corresponding aggregate values should be flagged for further investigation. Minor exceptions to allow for measurement system noise may be permitted in cases where the individual value is a large percentage of the aggregate value.

The deterministic checks listed above are based on theoretical relationships between the parameters. Empirical relationships can often be developed by reviewing historical data and noting parameter behavior which seldom or never occurs. Parameter relationship checks based on historical data of this kind are discussed in Section 3.3.3.

3.1.4 Data Processing Procedures

Reference 1 identifies 67 procedures currently in use for detecting and, when possible, correcting errors as they occur in computer systems. A review of reference 1 reveals that several procedures fall within the categories of internal, historical, and parallel data consistency checks while others are peculiar to data processing.[1] Some of the latter techniques are:

1. Context and staged edits (e.g., a field edit for checking the data values against the field specifications for length, character set, and value range).

2. Addition of quality flags to items in a data base to condition processing to avoid a mismatch between the data quality and its use.

3. Redundancy in batches, files, and inputs to improve reliability.

4. Checks on data sequence (e.g., input data are checked for correct time sequence).

5. Editing by classification of category, class limits, normal limits, and trend limits. For example, the behavior of an individual data item can be compared to its previous behavior or to the aggregate of individuals in its group. Procedures of this type are included in Sections 3.2 and 3.3.

6. Parallel check calculations, useful when the same results can be obtained by two independent calculation procedures.

7. Built-in test data, verification tests, and diagnostics to provide a test environment without the risk of allowing an unchecked program access to real files.

8. On-line testing to exercise the fault detection logic.

9. Clearly defined organizational responsibilities to ensure that the correct data validation procedures are continually employed.

No single grand or ideal solution to the problem of error control within automated data processing exists. Each organization must analyze its data processing system and the possible errors peculiar to the system. However, all organizations should develop a data flow diagram which indicates the steps in data handling at which an error can occur. In general these steps can be classified as (1) data input, (2) data transmission, (3) data processing and (4) data output.

Principal sources of error in the data input stage include key-punch errors and the use of mislabeled computer files. One should always review the input for errors of this kind. This can be done conviently if the input data are included in the computer output in a format for easy review.

The principal result of transmission error is the loss or alteration of data. A simple way to check for transmission error is to transmit the data a second time and then pair the two data streams. Gaps and alterations will be immediately apparent unless the transmission error is systematic.

Processing errors are more difficult to characterize and to detect. They are usually caused by deficiencies in the computer

programs which manipulate the data files, perform mathematic calculations, and format the output results. A standard method of checking for processing errors is to make up a small, typical data set, perform the appropriate data manipulations and calculations by hand, and compare with the results from the data processing system. This procedure should provide a good check if the data processing errors are not related to the size of the data set. This procedure is also appropriate for checking the part of the data processing system which outputs the data.

A detailed discussion of possible data validation procedures for computer applications would require a report of size at least equal to this entire report and would be repetitive of information available in the pertinent literature, some of which is listed in reference 1. Because reference 1 contains sufficient detail for computer trained personnel to develop a data validation system peculiar to their own system, no attempt is made here to duplicate its material.

3.2 TESTS FOR INTERNAL CONSISTENCY

Internal consistency tests check for values in a data set which appear atypical when compared to values of the whole data set. Common anomalies of this type include unusually high or low values (outliers) and large differences in adjacent values. The following tests for internal consistency are listed in order of increasing statistical sophistication. These tests will not detect errors which alter all values of the data set by either an additive or multiplicative factor (e.g., an error in the use of the scale of a meter or recorder).

3.2.1 Data Plots

Data plotting (including strip chart records) is one of the most effective means of identifying possible data anomalies. However, plotting all data points may require considerable manual effort or computer time. Nevertheless, data plots will often identify unusual data that would not ordinarily be identified by other internal consistency tests.

Figures 3-1 and 3-2 show computer plots of data with two types of unusual values. In Figure 3-1, there is an unusually high value which would be identified by almost all of the test procedures described in Section 3.2. On the other hand, the anomalous 4AM value in Figure 3-2 is similar in magnitude to several other values recorded for August 25 and 26 and would not be detected by most of these tests. The large difference between the 4AM value and the neighboring values in the time sequence is immediately apparent from the data plot.

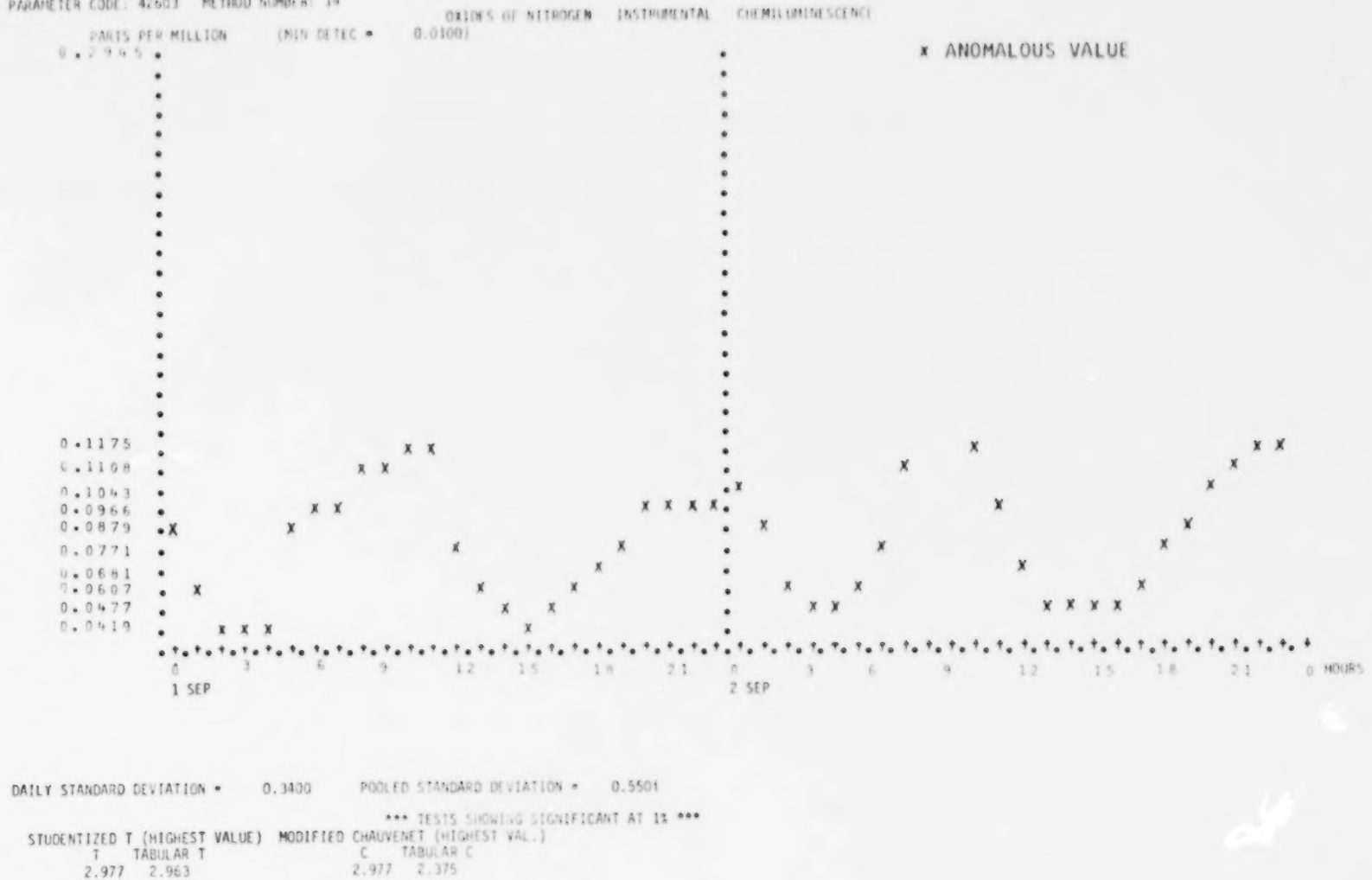
Although data plots are particularly appropriate for checking the internal consistency of data, they may also be used for checking historical consistency (e.g., the Shewhart control chart in Section 3.3.4) and parallel consistency (e.g., the intersite correlation test in Section 3.4.4).

3.2.2 Dixon Ratio Test

The Dixon ratio tests[3,4] are the simplest of the statistical tests recommended for evaluating the internal consistency of data. This section describes Dixon ratio tests for evaluating (1) the largest value in a data set and (2) the largest pair of values in a data set. Both procedures test the assumption H_o that the highest value(s) in a sample are consistent with the spread of the data. Other Dixon ratio tests which may be of interest to the data analyst are described in references 3 and 4.

3.2.2.1 Testing a Single High Value - The only data preparation required for testing a single high value is the identification of the lowest and the highest two or three (depending on n) values, where $x_1, x_2, x_3, \ldots, x_{n-2}, x_{n-1}, x_n$ is the arrangement of the lowest three and highest three values in ascending order of magnitude. The statistic of interest for $n \leq 7$ is the ratio of the difference between the highest and second highest values to the difference between the highest and lowest values (the range of the values in the data set); thus,

$$R = \frac{x_n - x_{n-1}}{x_n - x_1}. \qquad \text{Equation 3-1}$$



DAILY STANDARD DEVIATION = 0.3400 POOLED STANDARD DEVIATION = 0.5501

*** TESTS SHOWING SIGNIFICANT AT 1% ***

STUDENTIZED T (HIGHEST VALUE)		MODIFIED CHAUVENET (HIGHEST VAL.)	
T	TABULAR T	C	TABULAR C
2.977	2.963	2.977	2.375

Figure 3-1. Computer plot of data with single anomalous value.[2]

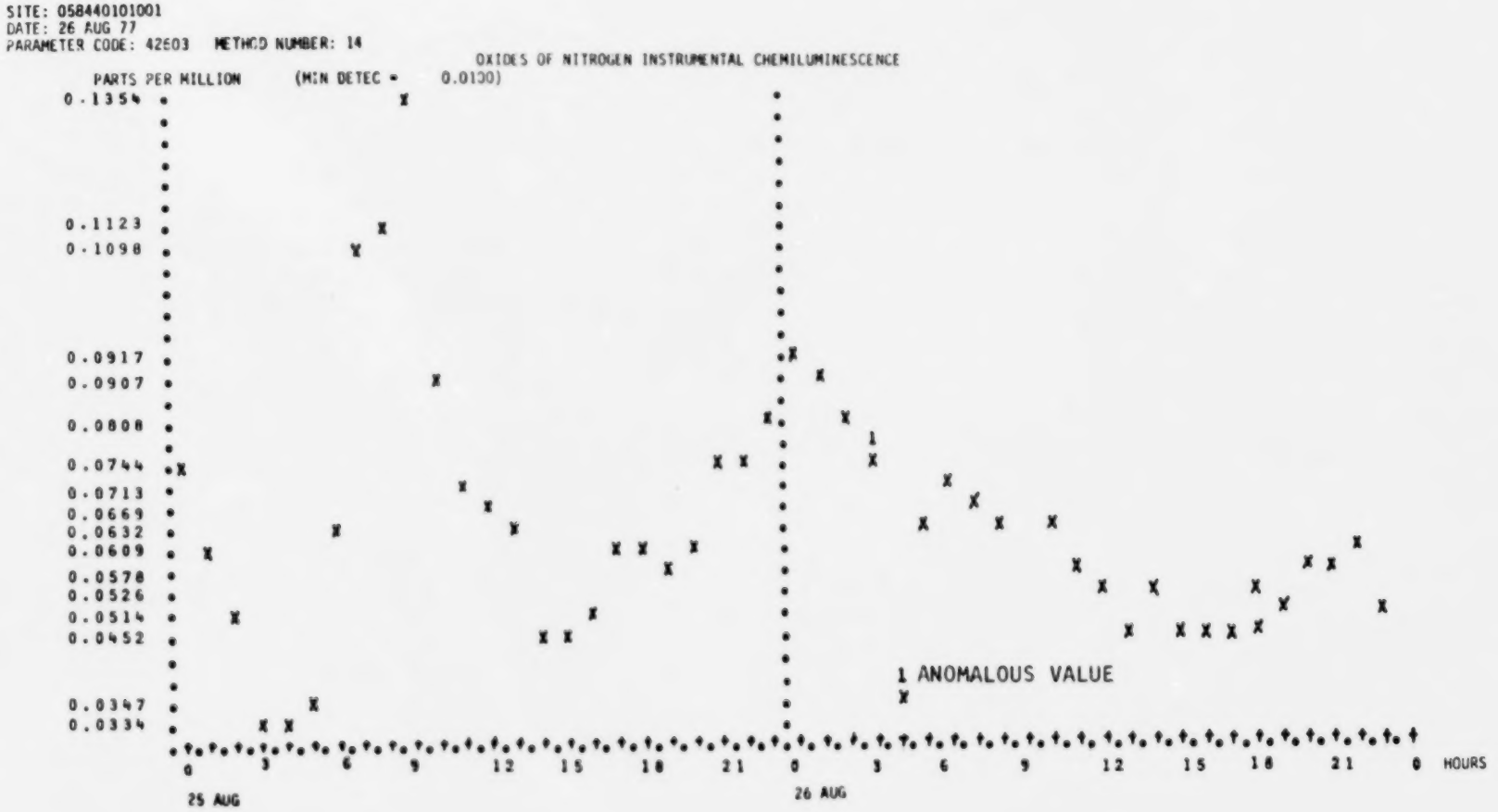


Figure 3-2. Computer plot of data with single anomalous value.[2]

The ratio R is a fraction between 0 and 1. As this ratio increases, the probability that the highest value is consistent with the rest of the data--that is, P[H_o is true]--decreases. Table A-1 (Appendix A) lists critical values of R associated with P[H_o is true] = 5% and P[H_o is true] = 1% for normally distributed data sets. Note that critical values are not listed for n values exceeding 25 and that the test procedure changes with n (Equation 3-1 is used for $n \leq 7$). The Dixon ratio test is not recommended for large data sets.

The above procedure assumes the data are normally and independently distributed. Since normal distributions are symmetrical in shape, they have equal mean and median values. However, air pollution data are not usually normally distributed. Data which are nonnormal are usually positively skewed (mean > median). In these cases a Dixon ratio test based on the lognormal distribution may be more appropriate; thus calculate R by using

$$R = \frac{\ln x_n - \ln x_{n-1}}{\ln x_n - \ln x_1}. \qquad \text{Equation 3-2}$$

Because Equation 3-2 is less sensitive to outliers of a normal distribution than Equation 3-1, it should be used only for testing data which are known to be adequately fitted by the lognormal distribution. If the data are not normally distributed and if one is unsure of the adequacy of the lognormal distribution, the gap test described in Section 3.2.4 should be considered.

Figure 3-3 illustrates the use of the Dixon ratio test in evaluating 2 months of 24-hour TSP data. The two data sets are identical except for the extreme values. Using the normality assumption (Equation 3-1), the Dixon ratio of data set A is

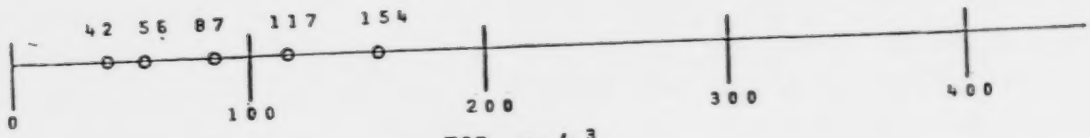
$$R_A = \frac{154 - 117}{154 - 42} = 0.33,$$

and the Dixon ratio of data set B is

$$R_B = \frac{420 - 154}{420 - 56} = 0.73.$$

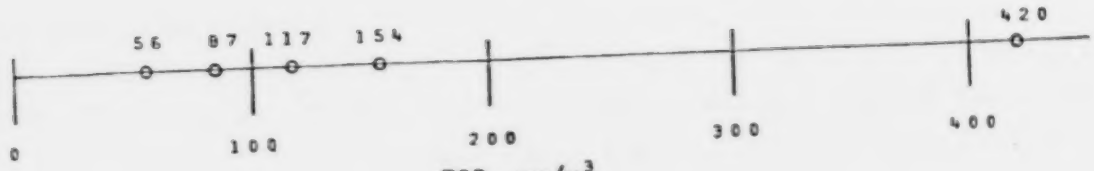
We accept the assumption H_o for data set A since 0.33 is smaller than 0.642, the 5% critical value for n = 5 in Table A-1. The value of 0.73 is larger than the 5% critical value but less than

DATA SET A



$$R_A = \frac{x_n - x_{n-1}}{x_n - x_1} = \frac{154-117}{154-42} = 0.33$$

DATA SET B



$$R_B = \frac{x_n - x_{n-1}}{x_n - x_1} = \frac{420-154}{420-56} = 0.73$$

Figure 3-3. Illustration of Dixon ratio test for two TSP monthly data sets.

the 1% critical value; consequently, $0.01 < P[H_o \text{ is true}] < 0.05$ and the value 420 appears to be inconsistent with the rest of the data set. Table A-1 contains several forms of the Dixon ratio test procedure as a function of sample size (n = 3 to 25) and whether the largest or smallest value is suspect. The application of the Dixon ratio test to air pollution data and the comparison of the test with other test procedures is given in references 5, 6, and 7.

3.2.2.2 Testing a Pair of High Values - The Dixon ratio test described above (Section 3.2.2.1) will not identify data sets which have two or more outliers of similar magnitude. To test for a pair of high values, calculate

$$R = \frac{x_n - x_{n-2}}{x_n - x_1} \qquad \text{Equation 3-3}$$

for normal data sets and

$$R = \frac{\ln x_n - \ln x_{n-2}}{\ln x_n - \ln x_1} \qquad \text{Equation 3-4}$$

for lognormal data sets. Decisions concerning acceptance or rejection of the assumption H_o--that the two highest values are consistent with the rest of the data set--are made according to the procedure described in Section 3.2.2.1 using Table A-2. Reference 3 describes related procedures appropriate for testing three or more outliers. Note: The Dixon ratio tests and similar statistical procedures for identifying outliers in a data set are not recommended for repeated use on the same data set. The error risks specified in tables were theoretically derived on the assumption that no extreme values have been removed from the data prior to the test. In practice, however, the tests are often applied successively. The user should be very cautions in doing this, that is, never discarding the data until a just cause has been determined. The user might also limit the rejected data to a small percentage of the data set, say 5%. In all applications, the treatment of the outliers should be documented in order that a subsequent user of the results will know the impact

of the outliers. For example, analyses can be given with and without the discarded data.

3.2.3 Grubbs Test

Like the Dixon ratio test described in Section 3.2.2.1, the Grubbs test[3,8] can be used to determine whether the largest observation x_n in a sample from a normal distribution is too large with respect to the internal consistency of the data set. The Grubbs test differs from the Dixon ratio test in that its test statistic,

$$T = \frac{x_n - \bar{x}}{s}, \qquad \text{Equation 3-5}$$

is calculated using all of the values in the data set. In addition to x_n, the data analyst must determine the arithmetic mean,

$$\bar{x} = \frac{1}{n} \Sigma\, x_i, \qquad \text{Equation 3-6}$$

and the standard deviation

$$s = [\Sigma(x_i - \bar{x})^2/(n - 1)]^{1/2}. \qquad \text{Equation 3-7}$$

The following equations can be used for positively skewed data which approximate the lognormal distribution:

$$\bar{x}' = \frac{1}{n} \Sigma \ln x_i \qquad \text{Equation 3-8}$$

$$s' = [\Sigma(\ln x_i - \bar{x}')^2/(n-1)]^{1/2}. \qquad \text{Equation 3-9}$$

High values of T indicate the likelihood that the maximum value is too high. Table A-3 gives upper probability levels for T for $3 \leq n \leq 147$.

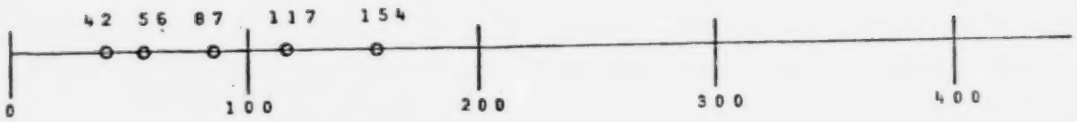
Figure 3-4 illustrates the Grubbs test in evaluating the 2 months of TSP data used in Section 3.2.2 to illustrate the Dixon ratio test. The T statistic for data set A is

$$T_A = \frac{154 - 91.2}{45.5} = 1.380.$$

Examining the n = 5 row in Table A-3, we find that $P[H_o \text{ is true}] > 10\%$ since $1.380 < 1.602$. It follows that a maximum value of 154 is not unexpected for this particular data set. The T statistic for data set B is

$$T_B = \frac{420 - 166.8}{146.1} = 1.733.$$

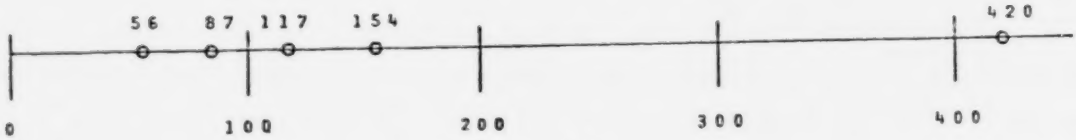
DATA SET A



TSP, $\mu g/m^3$

$$T_A = \frac{x_n - \bar{x}}{s} = \frac{154-91.2}{45.5} = 1.38$$

DATA SET B



TSP, $\mu g/m^3$

$$T_B = \frac{x_n - \bar{x}}{s} = \frac{420-166.8}{146.1} = 1.73$$

Figure 3-4. Illustration of Grubbs test for two TSP monthly data sets.

Table A-3 indicates that P[H_o is true] < 0.025, since 1.733 > 1.715. The value 420 appears to be inconsistent with the data set and should be flagged for further investigation.

3.2.4 Gap Test

The gap test[4,9] identifies spurious outliers by examining the frequency distribution for large gaps. The length of the gap between the largest x_n and the next largest value x_{n-1} is $x_n - x_{n-1}$, between the second and third largest values is $x_{n-1} - x_{n-2}$, and similarly for other gaps. The occurrence of a gap length larger than a predetermined critical value indicates a possible data anomaly.

The test assumes that the upper percentiles of the frequency distribution can be closely fit by a two-parameter exponential curve defined as

$$F(x) = 1 - \exp[-\lambda(x - \theta)]^* \qquad \text{Equation 3-10}$$

where F(x) is the fraction of the total observations less than or equal to x; λ is the slope parameter; and θ is the location parameter. Only the λ value of the fitted curve is used in the gap test. Values of λ can be estimated from the expression

$$\lambda = \frac{\ln [1 - F(x_1)] - \ln [1 - F(x_2)]}{x_2 - x_1} \qquad \text{Equation 3-11}$$

where $F(x_1)$ and $F(x_2)$ are the quantile values corresponding to the concentration values x_1 and x_2. For fits using three or more quantiles, use the least squares procedure for the 2-parameter exponential distribution described in Appendix B.

The probability of a gap of at least k units occurring in a data set free from erroneous values is

$$P[\text{gap} \geq k] = e^{-\lambda k}. \qquad \text{Equation 3-12}$$

A small P value indicates that the gap size is in question and that the corresponding data value(s) should be flagged.

Figure 3-5 illustrates the use of the gap test for hourly ozone data having a suspect value of 23 pphm. Substituting the

*exp [a] is a convenient means of writing (typing) e^a where e is the base for natural logarithms.

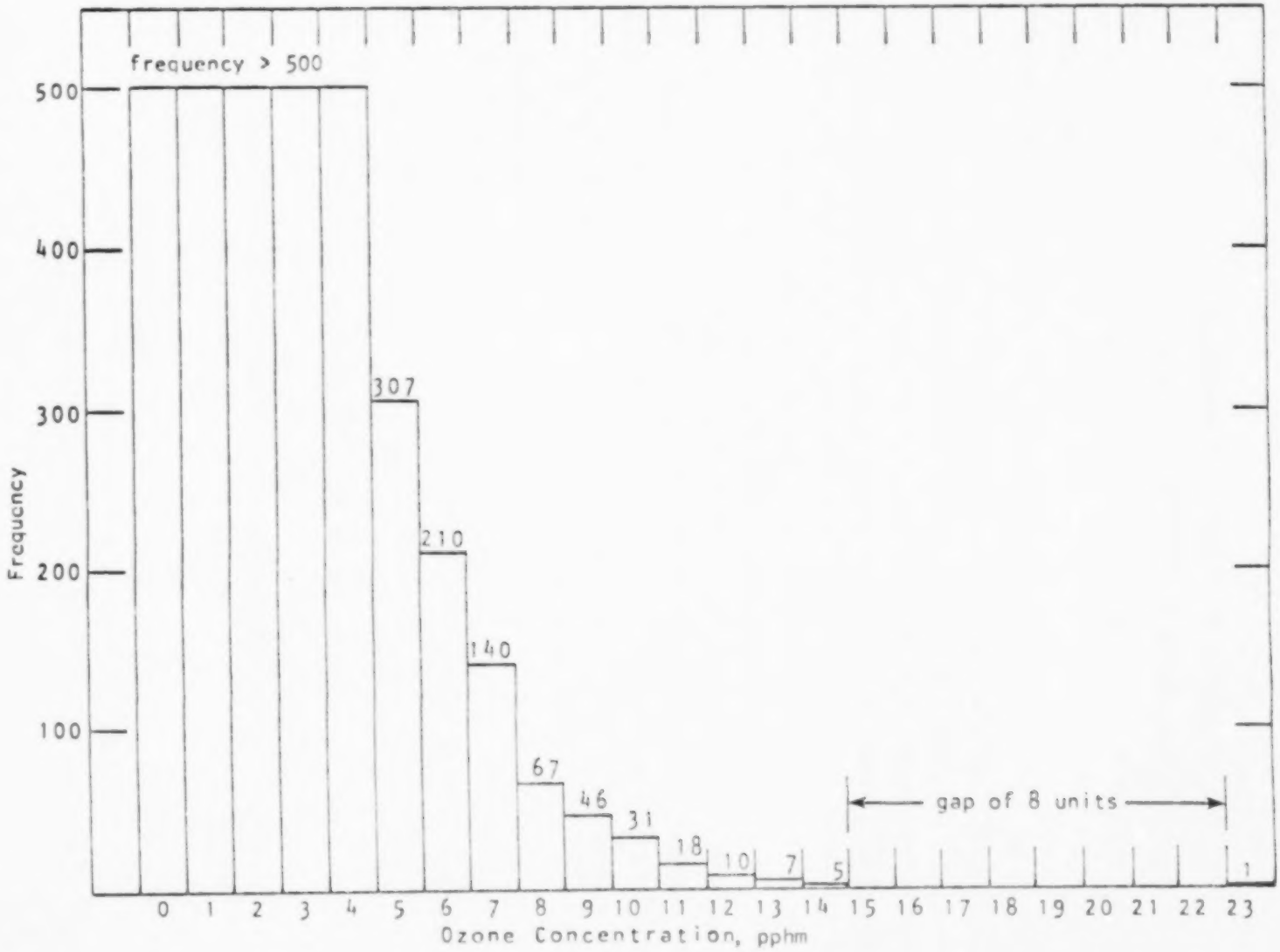


Figure 3-5. Illustration of gap test for hourly ozone data (Newport, KY, 1976).

90th and 99th percentile values, 4.5 pphm and 9.0 pphm, into Equation 3-11 yields

$$\lambda = \frac{\ln 0.10 - \ln 0.01}{9.0 - 4.5} = 0.512.$$

The gap is eight units in length; consequently

$$P[\text{gap} \geq 8] = e^{-(0.512)(8)} = 0.017.$$

There is less than a 2% probability that a gap of this length or larger would occur in the data set under the assumption stated. The 23-pphm value should be flagged for further investigation.

3.2.5 "Johnson" p Test

The "Johnson" p test assumes that the sampling distribution of n observations can be approximated by the cumulative distribution F(x), where F(x) is an identifiable function which defines the fraction of individuals in the sampled population having a value less than or equal to x. Let x_n be the largest recorded value in a sample of n observations from the population distribution F(x). The probability that all n observations from F(x) are less than x_n is $[F(x_n)]^n$; that is, the chance that a single observation is less than x_n, $F(x_n)$, raised to the power n. Thus the probability $P(x_n)$ that at least one value exceeds x_n (i.e., the largest value is at least x_n) is

$$P(x_n) = 1 - [F(x_n)]^n. \qquad \text{Equation 3-13}$$

We can now define an acceptance region for x_n by requiring that $P(x_n)$ fall between 0.05 and 0.95; that is,

$$0.05 < P(x_n) < 0.95. \qquad \text{Equation 3-14}$$

Data sets with maximum values which cause $P(x_n)$ to fall outside of the acceptance region should be flagged for further investigation.

The selection of an appropriate cumulative distribution F(x) to characterize the data is important in determining a reasonable acceptance region. Two distributions which often provide close fits to ambient air quality data are the Weibull and the lognormal.[10,11] Appendix B contains procedures for fitting these distributions to data and for selecting the appropriate F(x) function.

Figure 3-6 shows the probability function of Equation 3-13 corresponding to a Weibull distribution fitted to 1976 NO_2 data from Essex, MD. The limits of the acceptance region corresponding to $0.05 < P(x_n) < 0.95$ can be determined directly from the graph by noting that $P(165) = 0.05$ and $P(125) = 0.95$. The value of 140 ppb corresponds to $P(x_n) = 0.50$ and is considered a likely maximum value for this data set. If the data set has a recorded maximum value less than 125 ppb or higher than 165 ppb, it should be investigated further.

If the observed (recorded) maximum value is consistent with the overall distribution of the data set, there is a 10% probability that the maximum value of a valid data set will fall outside this acceptance region. If, in the general case, the data analyst decides to check all sets with x_n values outside the acceptance region, he or she would expect to check 1 out of 10 data sets unnecessarily. Where data validation requirements are more stringent, the acceptance range could be reduced to $0.10 < P(x_n) < 0.90$. On the other hand if valid data are being checked too frequently the acceptance range can be changed to $0.01 < P(x_n) < 0.99$. The acceptable range of x_n values should not be too narrow, since the P values are calculated from fitted distributions that only approximate the data set.

3.2.6 Multivariate Tests

The procedures given in Section 3.2.2 through 3.2.5 can be used for testing data sets involving more than one variable by applying them independently to each variable; however, this approach may be inefficient, particularly when the variables are statistically correlated. In this case, the data analyst should consider the test procedure given in Section 3.4.4 for correlated data. Although this test is included under parallel data sets because of the particular application to correlated data from two sets, it can be applied to correlated variables within the same data set; for example, to the concentrations of two pollutants, such as TSP and Pb. In some cases, a multivariate test of this kind will show that a value of one variable that appears to be an

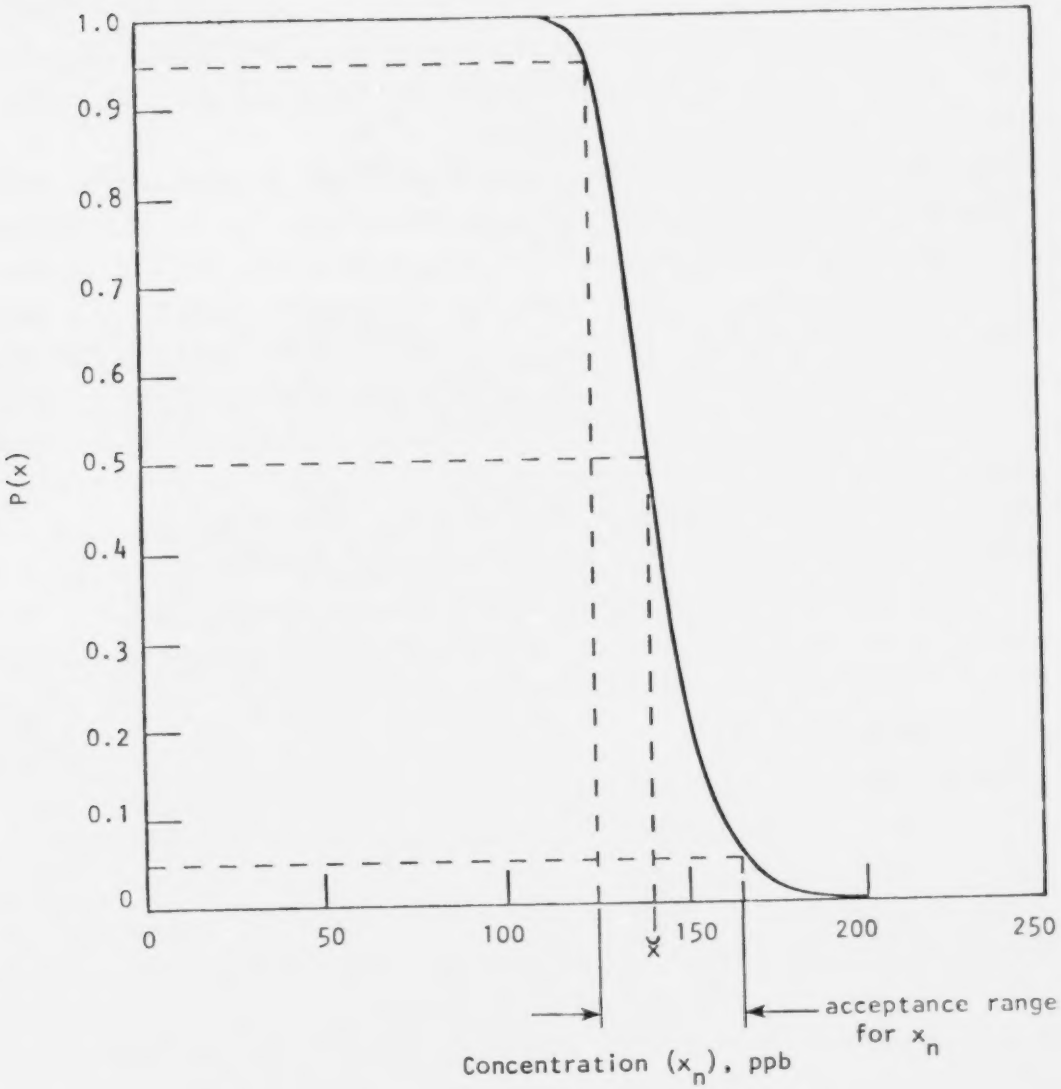


Figure 3-6. P(x) versus concentration for Weibull distribution fitted to 1976 NO_2 data from Essex, MD. Acceptance range defined as $0.05 < P(x) < 0.95$.

outlier using a single variable test procedure will be consistent with the data set when one or more other variables are considered. Conversely, there may be a value of one variable that is consistent with the other data in the data set when only one variable is considered, but is definitely a possible outlier when two variables are considered.

Multivariate test procedures which have been successfully used to perform data validation checks include cluster analysis,[12] principal component analysis,[13] and correlation analyses. Applications of these usually require computerized procedures. For example, cluster analysis can be applied using a program called NORMIX.[14]

3.3 TESTS FOR HISTORICAL CONSISTENCY

Tests for historical consistency check the data set with respect to similar data recorded in the past. It is important to note that some of the data validation procedures to be described in this section will detect relatively small changes in the average value and/or the dispersion of the data values. In particular, these procedures will detect changes where each item is increased (decreased) by a constant or by a multiplicative factor. This is not the case for the procedures in Section 3.2 which yield the same value for the test statistic when all data are changed by the same constant or multiplicative factor.

3.3.1 Gross Limit Checks

Gross limit checks are useful in detecting data values that are either highly unlikely or generally considered impossible. Upper and lower limits are developed by examining historical data for a site (or for other sites in the area). Whenever possible, the limits should be specific to each monitoring site and should consider both the parameter and instrument/method characteristics. Table 3-1 shows examples of gross limit checks that have been used for ambient air monitoring data in the St. Louis area.[15,16] Although this technique can be easily adapted to computer application, it is particularly appropriate for technicians who reduce data manually or who scan strip charts to detect unusual events.

TABLE 3-1. EXAMPLES OF HOURLY AND DAILY GROSS LIMIT CHECKS FOR AMBIENT AIR MONITORING[15,16]

Parameter	Limits	
	Lower	Upper
O_3	0 ppm	1 ppm[a]
NO_2	0 ppm	2 ppm
NO	0 ppm	3 ppm
NO_x	0 ppm	5 ppm
Total suspended particulates	0 $\mu g/m^3$	2000 $\mu g/m^3$[b]
CO	0 ppm	100 ppm
Total hydrocarbons	0 ppm	25 ppm[a]
Methane	0 ppm	25 ppm[a]
Total sulfur	0 ppm	1 ppm
SO_2	0 ppm	1 ppm
H_2S	0 ppm	1 ppm
Aerosol scatter	0.000001 m^{-1}	0.0040 m^{-1}[a]
Windspeed	0 m/s	22.2 m/s
Wind direction	0°	360° (540° for some wind systems)[a]
Temperature	-20°C	45°C
Dewpoint	-30°C	45°C
Temperature gradient	-5°C	5°C
Barometric pressure	950 mb	1050 mb

[a]These limits have been changed from the original in reference 15 and 16 based on after-the-fact considerations.

[b]Upper limit for a 24-hour average.

3.3.2 Pattern and Successive Value Tests

The pattern and successive value tests check the data for pollutant behavior which has never or very rarely occurred. Like the gross limit checks, they require that a set of boundary values or limits be determined empirically from prescreened historical data. Values representing pollutant behavior outside of these predetermined limits are then flagged for further investigation.

EPA has recommended the use of the pattern tests which place upper limits on:

1. The individual concentration value (maximum hour test),
2. The difference in adjacent concentration values (adjacent hour test),
3. The difference or percentage difference between a value and both of its adjacent values (spike test), and
4. The average of four or more consecutive values (consecutive value test).[4]

The maximum hour test (a form of gross limit check) can be used with both continuous and intermittent data. The other three tests should be used only with continuous data.

Table 3-2 is a summary of limit values developed by EPA for hourly average data. These values were selected on the basis of empirical tests on actual data sets. Note that the limit values vary with different data stratifications (e.g., day/night).

These limit values will usually be inappropriate for other pollutants, data stratifications, averaging times, or EPA regions. In these cases, the data analyst should develop the required limit values by examining historical data similar to the data being tested. These limit values can be later modified if they flag too many values that are later proven to be correct or if they flag too few errors. Pattern tests should continue to evolve to meet the needs of the analyst and the characteristics of the data.

3.3.3 Parameter Relationship Test

Parameter relationship tests can be divided into two main categories: deterministic tests involving the theoretical rela-

TABLE 3.2. SUMMARY OF LIMIT VALUES USED IN EPA REGION V FOR PATTERNS TESTS[4]

Pollutant (units)	Data stratification	Maximum hour	Adjacent hour	Spike	Consecutive 4-hour
Ozone-total oxidant ($\mu g/m^3$)	summer day	1000	300	200(300%)	500
	summer night	750	200	100(300%)	500
	winter day	500	250	200(300%)	500
	winter night	300	200	100(300%)	300
Carbon monoxide (mg/m^3)	rush traffic hours	75	25	20(500%)	40
	nonrush traffic hours	50	25	20(500%)	40
Sulfur dioxide ($\mu g/m^3$)	EPA region	2600	500	200(500%)	1000
Nitrogen dioxide ($\mu g/m^3$)	None	1200	500	200(300%)	1000

tionships between parameters (e.g., NO < NO_x) and empirical tests which check whether or not a parameter is behaving normally in relation to the observed behavior of one or more other parameters. Deterministic parameter checks are discussed in Section 3.1.3; empirical parameter checks are discussed in this section since determining the "normal" behavior of related parameters requires the detailed review of historical data.

The following area-specific example illustrates the testing of meteorological data using a combination of successive value tests, gross limit tests, and parameter relationship tests. One should consult with the local National Weather Service office for relationships for a specific area. The validation protocol calls for the following procedures to be applied to ambient temperature data based on the availability of hourly averages reported in monthly formats:

1. Check the hourly average temperature. The minimum should occur between 04-09 hours, and the maximum should occur between 12-17 hours.

2. Inspect the hourly data for each day. Hourly changes should not exceed 10°F. If a decrease of 10°F or more occurs, check the wind direction and the precipitation summaries. The wind direction should have changed to a more northerly direction, and/or rainfall of 0.15 in. or more per hour should have fallen.

3. Hourly values should not exceed predetermined maximum or minimum values based on month of the year. For example, in November the maximum allowable temperature is 85°F and the minimum allowable temperature is 10°F.

If any of the above criteria are not met, then the data for the appropriate time period are flagged for anomaly investigation.

In this example, relationship checks have been developed for temperature and wind direction as well as temperature and precipitation. Other pairs of parameters for which relationship checks could be developed include solar insolation and cloud cover; windspeed aloft and ground windspeed; ozone and NO; and temperature and humidity.

3.3.4 Shewhart Control Chart

The gross limit checks and the pattern tests described in Sections 3.3.1 and 3.3.2, respectively, use critical values ("control limits") based on historical data to identify possible data anomalies involving single values or small numbers of consecutive values. The Shewhart control chart is a valuable supplement to these two tests in that it identifies data sets which have mean or range values that are inconsistent with past data sets.

The classical use of a quality control chart is to determine the limits on the basis of historical data and to apply these limits to future data to determine the state of control. In the data validation process the control chart can be used in this classical sense (particularly at the local agency) or in an after-the-fact sense. In the latter case, the control chart technique may be applied to data already recorded/documented by using a portion of the data to determine the limits to be used for the remaining data. The discussion of quality control charts herein applies to either use (classical or after-the-fact).

The first step in the development of a Shewhart control chart is the selection of a suitable sample size for the test. Each sample should contain between 4 and 15 values and should represent a well-defined time period (day, month, quarter, etc.) for which there is at least 10 to 15 historical data samples. Where possible, these time periods should relate to the NAAQS (National Ambient Air Quality Standards) of interest. Months or quarters would be appropriate time periods for tests of 24-hour TSP, SO_2, and NO_2 data collected at 6- or 12-day intervals.

The second step is to calculate the limits on the Shewhart control chart following the procedures in Appendix C or in a standard reference text.[18] The details of the calculation will be illustrated by the following example.

Average 24-hour TSP concentrations at Philadelphia monitoring site 397140014H01 are recorded every sixth day. We desire to use a Shewhart control chart to check 1978 data as they are reported. A sample size of five is chosen so that incoming data will be checked 12 times a year if all values are recorded. A review of the TSP data for 1975-77 reveals 25 months which contain exactly five TSP values. Table 3-3 lists the mean and the range of each of these data sets. No seasonal trends are apparent over the 3-year period. We apply the Grubbs test (Section 3.2.3) to the data sets with suspicious range values (>100). In each case, $T < 1.602$, $P[H_o \text{ is true}] > 10\%$, so we decide not to reject the data set. These 25 data sets now form the historical data base for determining the limits for the control chart.

Since all data sets, including the one to be tested, contain five observations, Equations C-1, C-5, C-6, and C-10 (Appendix C) are used with $d_2 = 2.326$ and $c_2 = 0.8407$. If $z = 2$, then

$\bar{\bar{x}}$ (mean of $\bar{x}$'s) = 56.5

$\sigma_{\bar{x}}$ (standard deviation of the mean) = 9.0

$UL_{\bar{x}}$ (upper 2σ limit for the mean) = 74.5

$LL_{\bar{x}}$ (lower 2σ limit for the mean) = 38.5

$\bar{R}$ (mean range) = 47.0

UL_R (upper 2σ limit for R) = 80.9

LL_R (lower 2σ limit for R) = 13.0

TABLE 3-3. TSP DATA FROM SITE 397140014H01 SELECTED AS HISTORICAL DATA BASE FOR SHEWHART CONTROL CHART (1975-1977)

Month-year	Mean ($\bar{x}$), $\mu g/m^3$	Range (R), $\mu g/m^3$	Month-year	Mean ($\bar{x}$), $\mu g/m^3$	Range (R), $\mu g/m^3$
1-75	54.6	67	10-76	34.6	50
5-75	63.8	39	11-76	53.4	29
6-75	59.0	25	12-76	52.2	44
7-75	63.0	23	3-77	40.4	28
8-75	68.2	54	4-77	63.6	57
10-75	41.8	26	6-77	45.4	31
11-75	68.4	81	7-77	53.4	19
12-75	57.6	39	8-77	58.6	26
1-76	82.4	87	9-77	46.0	12
4-76	90.2	117	10-77	45.6	33
5-76	43.8	48	11-77	49.8	54
7-76	72.6	80	12-77	30.4	22
9-76	73.4	83			

Figures 3-7 and 3-8 have been constructed using these values, and the mean and range values from eleven 1978 data sets have been plotted. The raw data for 1978 are in Table 3-4.

TABLE 3-4. TSP DATA FROM SITE 397140014H01 FOR CONTROL CHART (1978)

Data set	Month	Mean	Range
1	1	30.6	27
2	2	47.4	60
3	3	54.4	39
4	4	31.8	29
5	5	53.6	46
6	6	64.8	46
7	8	68.8	87
8	9	43.2	31
9	10	52.4	59
10	11	60.8	71
11	12	31.6	22

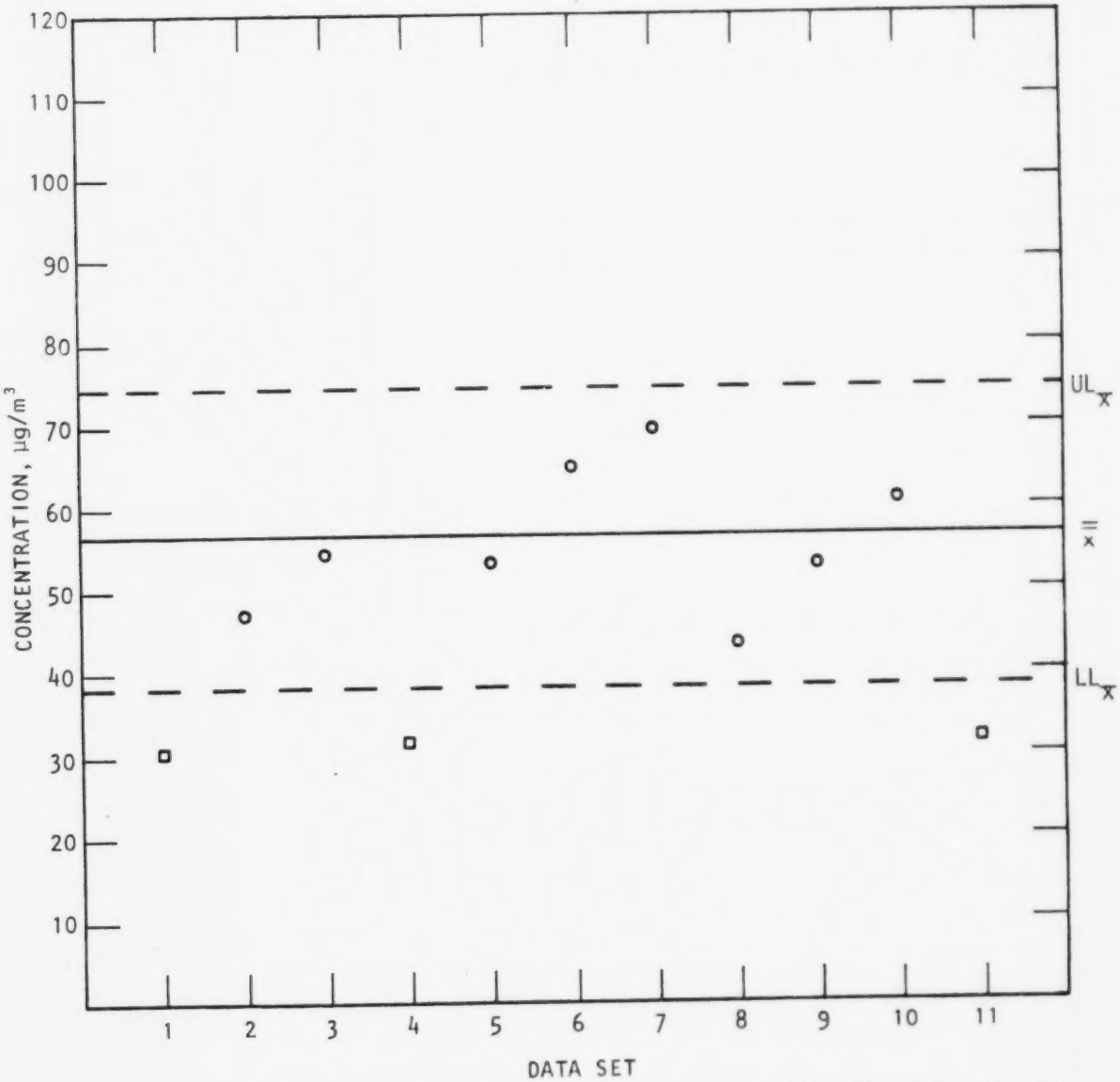


Figure 3-7. Shewhart control chart for mean values with 1978 data plotted.

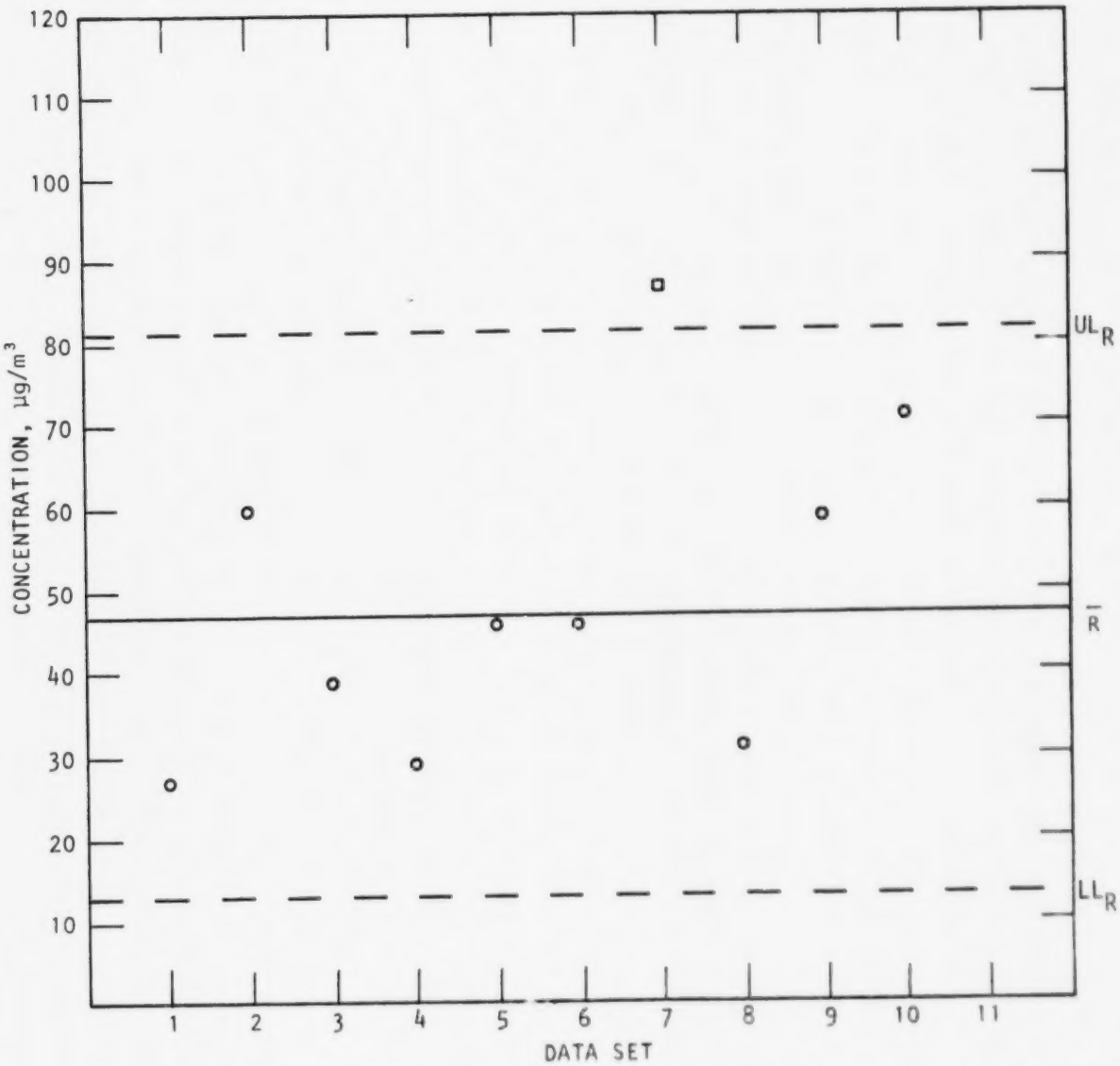


Figure 3-8. Shewhart control chart for range values with 1978 data plotted.

Figure 3-7 shows three mean values below the $LL_{\bar{x}}$, but no mean values above the $UL_{\bar{x}}$. The overall distribution of plotted points suggests a possible explanation for the anomalous low values. Of the 11 points, eight are below the $\bar{\bar{x}}$ line while only three are above it. We should investigate two hypotheses: (1) air quality has improved and (2) the TSP monitor has developed a negative bias. The first hypothesis can be checked by seeing if TSP data from nearby monitors show similar trends. Measurement bias may be revealed through calibration checks and careful inspection of quality assurance records. In particular, check for changes in the location of the sampling mechanism.

The 1978 range values also tend to be smaller than expected; seven are less than $\bar{R}$ while only four are greater than $\bar{R}$ (Figure 3-8). None of the range values are less than the $LL_{\bar{R}}$, however. The sole outlier is slightly above the $UL_{\bar{R}}$. A good preliminary check of data sets with anomalous range vaues is the Grubbs test. Following the procedure described in Section 3.2.3, we find that $x_n = 129$, $\bar{x} = 68.8$, $s = 34.6$, and $T = 1.740$. Since inspection of Table A-3 reveals that $0.010 < P[H_o \text{ is true}] < 0.025$ when $T = 1.740$, we are suspicious of the highest value, $x_n = 129$. Further investigation should focus on determining the validity of this measurement.

Assuming that we have determined the cause(s) of the anomalies noted on Figures 3-7 and 3-8, we must now decide on the appropriate control limits for plotting 1979 data. If an improvement in air quality has indeed occurred, we should calculate new control limits which reflect the lower $\bar{x}$ values expected. A recommended procedure in this case is to omit the earliest year of data and to add the most recent. If the problem is traced to measurement bias, either the original control limits can be retained or new control limits can be calculated by adding any valid 1978 data sets to the 25 data sets in the historical data base.

It should be apparent from this example that a physical control chart need not be constructed to test individual mean and range values; a computer program can be easily developed which

will list all data sets which fail the test. The main benefit of the visible control chart is its capability to indicate trends and other data patterns. If new data are consistent with historical data, the points plotted on the control chart should fall within the limits, randomly above and below the central line. A long run of points on one side of the line (even though none of the points lies outside the control limits) may indicate a systematic bias in the data or an actual trend in air quality that requires investigation. Common practice is to check these possibilities whenever the run exceeds 6 points. Used in this way, the control chart can warn the data analyst of data problems before they become serious.

3.4 TESTS FOR CONSISTENCY OF PARALLEL DATA SETS

The tests for internal consistency described in Section 3.2 implicitly assume that most of the values in a data set are correct. Consequently, if all of the values in a data set incorporate a small positive bias, tests such as the Dixon ratio test would not identify the data set as inconsistent. One method of identifying a systematic bias of this type is to compare the data set with other data sets which presumably have been sampled from the same population (i.e., same air mass and time period) and to check for differences in the average value or the overall distribution of values. This section describes four such procedures--the sign test, the Wilcoxon signed-rank test, the Wilcoxon rank sum test, and the intersite correlation test--which are recommended for comparisons involving two "parallel" data sets. These four tests are presented in order of increasing sensitivity to differences between the data sets and increasing computational complexity. The first three tests are nonparametric; that is, they do not assume that the data have a particular distribution.[19] Consequently, these tests can be used for the nonnormal data sets which frequently occur in air quality analysis.

3.4.1 The Sign Test

The sign test[19] is a relatively simple procedure for testing the assumption H_o--that two related (paired) samples, such as

data sets from adjacent monitoring instruments, have the same median. The data analyst simply determines the sign (+ or -) of the algebraic difference between each of the pairs of data points, and then counts the total number of positive signs (n_+) and negative signs (n_-); differences of zero are ignored. The probability that both samples have the same median is

$$P[H_o \text{ is true}] = 2\, N!(1/2)^N \sum_{j=0}^{n} \frac{1}{j!(N-j)!} \qquad \text{Equation 3-15}$$

where n is the smaller of the two numbers n_+ and n_-, and $N = n_+ + n_-$. For large N (say >25), one can use the normal approximation to the above probability by calculating

$$z = \frac{2n - N}{\sqrt{N}}. \qquad \text{Equation 3-16}$$

In this case, $P[H_o \text{ is true}]$ is equal to twice the area to the left of z under the standard normal curve. The following table lists P values corresponding to selected values of z.

z	P	z	P
-1.282	0.20	-2.326	0.020
-1.645	0.10	-2.576	0.010
-1.960	0.05	-2.807	0.005

A z value of -1.85 would imply that the probability of the samples representing populations with the same distribution is between 5% and 10%.

Table 3-5 lists ozone data recorded August 1, 1978, at two monitoring stations in Washington, D.C. The difference column contains 13 positive values and 6 negative values. Consequently, $n_+ = 13$, $n_- = 6$, n = 6, and N = 19. Substituting these values in Equation 3-15 yields

$$P[H_o \text{ is true}] = (2)(19!)(1/2)^{19} \sum_{j=0}^{6} \frac{1}{j!(19-j)!} = 0.167.$$

If the normal approximation (Equation 3-16) is used, we have

$$z = \frac{12 - 19}{\sqrt{19}} = -1.606,$$

and $0.10 < P[H_o \text{ is true}] < 0.20$. Generally the data analyst will not reject H_o if $P[H_o \text{ is true}] > 0.05$.

TABLE 3-5. OZONE DATA (ppb) RECORDED AUGUST 1, 1978, AT MONITORING SITES 090020011I01 (SITE A) AND 090020013I01 (SITE B)

Hour	Site A	Site B	Difference
1	65	50	+15
2	40	50	-10
3	35	45	-10
4	30	35	-5
5	30	25	+5
6	15	15	0
7	15	10	+5
8	5	5	0
9	10	5	+5
10	10	10	0
11	35	50	-15
12	65	55	+10
13	65	60	+5
14	100	90	+10
15	130	110	+20
16	90	65	+25
17	70	65	+5
18	70	70	0
19	85	65	+20
20	55	50	+5
21	45	35	+10
22	25	40	-15
23	20	20	0
24	20	30	-10

Table 3-6 was developed to simulate a calibration error in the site A data. Each reading in the site A column in Table 3-5 was increased by 5 ppb. Readings in the site B column were not changed. There are now 18 positive values and 5 negative values. It follows that $n_+ = 18$, $n_- = 5$, $n = 5$, and $N = 23$. Using Equation 3-15, the probability that H_o is true is

$$P[H_o \text{ is true}] = (2)(23!)(1/2)^{23} \sum_{j=0}^{5} \frac{1}{j!(23-j)!} = 0.0106.$$

The normal approximation (Equation 3-16) yields

$$z = \frac{2(5) - 23}{\sqrt{23}} = -2.711,$$

corresponding to $P[H_o \text{ is true}] = 0.007$. In either case, $P[H_o \text{ is true}] < 0.05$, and there is good reason to believe that a fundamental inconsistency exists between the two data sets.

3.4.2 Wilcoxon Signed-Rank Test

Like the sign test, the signed-rank test can be used to test the assumption (H_o) that two samples come from populations having the same medians. The Wilcoxon test is generally more powerful than the sign test since it considers both the sign and the magnitude of the difference between paired data. Table 3-7 lists the steps in the procedure. The large sample approximation given in step 5 is appropriate when $N > 20$.

Table 3-8 illustrates the application of the signed-rank test to the ozone data listed in Table 3-5. The absolute differences were first listed according to increasing magnitude and then assigned ranks. Note that differences of zero are ignored; there are 19 nonzero differences. The sum of the ranks associated with the minus sign (T_1) is 65.5 and the sum of the ranks associated with the plus sign (T_2) is 124.5. Table A-4 indicates that when $N = 19$ the probability of H_o being true would be less than 10% if $T_2 \leq 53$ or $T_2 \geq 137$. Since $53 < 124.5 < 137$, we can say $P[H_o \text{ is true}] > 0.10$. Using the large sample approximation (Equation 3-17),

$$z = \frac{T_1 - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}},$$ Equation 3-17

we find that

$$z = \frac{65.5 - \frac{19(19+1)}{4}}{\sqrt{\frac{(19)(19+1)[(2)(19)+1]}{24}}} = -1.187,$$

which corresponds to $P[H_o \text{ is true}] = 0.235$. In neither case is $P[H_o \text{ is true}] < 0.10$; consequently, there is not good reason to reject H_o.

TABLE 3-6. OZONE DATA (ppb) INCORPORATING SIMULATED +5 ppb CALIBRATION ERROR AT SITE A

Hour	Site A	Site B	Difference
1	70	50	+20
2	45	50	-5
3	40	45	-5
4	35	35	0
5	35	25	+10
6	20	15	+5
7	20	10	+10
8	10	5	+5
9	15	5	+10
10	15	10	+5
11	40	50	-10
12	70	55	+15
13	70	60	+10
14	105	90	+15
15	135	110	+25
16	95	65	+30
17	75	65	+10
18	75	70	+5
19	90	65	+25
20	60	50	+10
21	50	35	+15
22	30	40	-10
23	25	20	+5
24	25	30	-5

TABLE 3-7. PROCEDURE FOR WILCOXON SIGNED-RANK TEST[19]

1. Determine the sign and magnitude of the algebraic difference between each of the pairs of data points; assume that N scores remain after throwing out all zeros.
2. Assign ranks to the absolute values of these N differences; use the average rank in case of ties. Make sure that the ranks increase as the absolute values increase.
3. Assign to each rank the sign which it represents.
4. Determine the sum of the ranks associated with a minus sign (T_1) and the sum of the ranks associated with a plus sign (T_2).
5. To test H_o, either use published tables (see Appendix A, Table A-4) based on T_2 and N or use a large sample normal approximation with

$$z = \frac{T_1 - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \quad \text{Equation 3-17}$$

The probability that H_o is true--that is, P[H_o is true]--is equal to twice the area to the left of z under the standard normal curve.

If we repeat the signed-rank test with the biased ozone data in Table 3-6, we find that $T_1 = 38.5$, $T_2 = 237.5$, and $N = 23$. Table A-4 indicates that if $T_2 \leq 54$ or if $T_2 \geq 222$ when $N = 23$, then P[H_o is true] < 0.01. Since 237.5 > 222 we reject H_o; there is good reason to believe that samples A and B represent populations with different medians. The normal approximation supports this conclusion since

$$z = \frac{38.5 + \frac{23(23+1)}{4}}{\sqrt{\frac{(23)(23+1)[(2)(23)+1]}{24}}} = -3.026$$

and a z value of -3.026 corresponds to P[H_o is true] = 0.0025.

3.4.3 Rank Sum Test

The rank sum procedure is useful in testing the assumption (H_o) that two samples represent populations with the same distribution. Unlike the sign test and the ranked sign test, the rank

and substituting the data,

$$z = \frac{598 - \frac{(24)(24+24+1)}{2}}{\sqrt{\frac{(24)(24)(24+24+1)}{12}}} = 0.206.$$

A z value of 0.206 corresponds to $P[H_o \text{ is true}] = 0.837$.

TABLE 3-10. APPLICATION OF RANK SUM TEST TO OZONE DATA IN TABLE 3-5

Value	Data set	Rank	Value	Data set	Rank
5	A	2	45	B	25.5
5	B	2	45	A	25.5
5	B	2	50	B	28.5
10	B	5.5	50	B	28.5
10	A	5.5	50	B	28.5
10	A	5.5	50	B	28.5
10	B	5.5	55	B	31.5
15	A	9	55	A	31.5
15	B	9	60	B	33
15	A	9	65	A	36.5
20	A	12	65	A	36.5
20	B	12	65	A	36.5
20	A	12	65	B	36.5
25	B	14.5	65	B	36.5
25	A	14.5	65	B	36.5
30	A	17	70	A	41
30	A	17	70	A	41
30	B	17	70	B	41
35	A	20.5	85	A	43
35	B	20.5	90	B	44.5
35	A	20.5	90	A	44.5
35	B	20.5	100	A	46
40	A	23.5	110	B	47
40	B	23.5	130	A	48

Application of the rank sum test to the biased data in Table 3-6 yields $T_1 = 626.5$. Table A-5 shows that $T_\ell \leq 626.5 \leq T_r$ for $\alpha = 0.10$; it follows that $P[H_o \text{ is true}] > 0.10$. The large sample normal approximation is consistent with this result since $z = 0.794$ and $P[H_o \text{ is true}] = 0.427$. In either case, the rank sum test does not indicate that H_o should be rejected.

Note that the rank sum test accepts H_o with respect to the data in Table 3-6, but the sign test and the signed-rank test reject H_o. This apparent contradiction is the result of a fundamental difference in the construction of the tests. The sign and signed-rank tests are paired-value tests; consequently, they are particularly sensitive to differences between related observations. Both of these tests changed from accepting H_o to rejecting H_o when a small bias (5 ppb) was introduced into data set A. Because the rank sum test compares sample distributions without actually pairing the data, it is relatively insensitive to small data shifts. The rank sum test is more appropriate for identifying censored samples.

The data listed in Table 3-10 are similar to those in Table 3-5, except that all values from site A exceeding 60 ppb have been omitted. In this case $n_1 = 15$, $n_2 = 24$, and $T_1 = 225$. The largest α value in Table A-5 for which $T_\ell \leq T_1 \leq T_r$ is 0.02. Consequently, $0.02 < P[H_o \text{ is true}] \leq 0.05$, and there is good reason to reject H_o. Substitution of the appropriate values into the equation for the large sample approximation yields $z = -2.165$ and $P[H_o \text{ is true}] = 0.030$.

The sign and signed-rank tests are not as sensitive to censored data as the rank sum test, mainly because they ignore unpaired data. In the example just described, the sign and signed-rank tests would not use the nine values from site B for which there are no corresponding values listed for site A. Unlike the rank sum test, neither of these paired data tests would have rejected H_o.

3.4.4 Intersite Correlation Test

The intersite correlation test is suggested as a means of comparing two correlated parameters being measured at the same

TABLE 3-8. APPLICATION OF WILCOXON SIGNED-RANK TEST TO OZONE DATA (ppb) IN TABLE 3-4

Absolute difference	(Signed) Rank
0	*
0	*
0	*
0	*
0	*
5	-4
5	+4
5	+4
5	+4
5	+4
5	+4
5	+4
10	-10.5
10	-10.5
10	+10.5
10	+10.5
10	+10.5
10	-10.5
15	+15
15	-15
15	-15
20	+17.5
20	+17.5
25	+19

sum test is applicable to independent (unrelated) samples. Table 3-9 lists the steps in the procedure appropriate for tests of 10 or more data pairs.

TABLE 3-9. PROCEDURE FOR WILCOXON RANK SUM TEST

1. Combine the n_1 observations from population 1 and the n_2 observations from population 2, arrange them in ascending order of size, and then assign ranks from 1 to $(n_1 + n_2)$ where $n_1 \leq n_2$. In case of ties, use the average rank.
2. Calculate T_1, the sum of the ranks assigned to observations from population 1.
3. Compare T_1 with the critical values in Table A-5, $\alpha = 0.10$.[20] If $T_\ell \leq T_1 \leq T_r$ for the values of T_ℓ and T_r listed for samples of n_1 and n_2 in the table, then the P[H_o is true] ≥ 0.10. If T_1 is outside the indicated range, go to the $\alpha = 0.05$ table and repeat test. If $T_\ell \leq T_1 \leq T_r$, then $0.05 \leq$ P[H_o is true] ≤ 0.10. If not, continue search until α value is found for which $T_\ell \leq T_1 \leq T_r$. If $\alpha < 0.05$ there is good reason to reject H_o.
4. To test H_o for $n_2 \geq 20$, use the large sample normal approximation, Equation 3-18. The probability that H_o is true, i.e., P[H_o is true], is equal to twice the area to the right of z under the standard normal curve.

Table 3-10 illustrates the application of the rank sum test to the ozone data listed in Table 3-5. Note that tied values are assigned an average rank value. The sum T_1 of the ranks from the first sample is equal to 598. In Table A-5, $\alpha = 0.10$, $T_\ell = 507$ and $T_r = 669$ for $n_1 = n_2 = 24$.[20] Consequently $T_\ell \leq T_1 \leq T_r$, P[H_o is true] ≥ 0.10; thus H_o should not be rejected. Using the large sample normal approximation, (Equation 3-18),

$$z = \frac{T_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \qquad \text{Equation 3-18}$$

site or at neighboring sites. An example is given to illustrate the procedure for TSP measurements on every sixth day at each of the neighboring sites over a period of 1 year.

It would be possible to treat each site independently; however, this approach does not consider the relationship between the measured TSP concentrations at the two sites, and hence may err either by not identifying a potential data anomaly or by identifying a possible outlier and later finding that it is consistent with that measured at a neighboring site.

The data for the example are in Table 3-11. Denote by x and y the measurements of μg TSP/m^3 at sites 397140020H01 and 397140014H01, respectively. The original data and the logarithms are both given because TSP data are almost always better approximated by using the lognormal distribution.

Because the lognormal distribution is preferable, these data are plotted in Figure 3-9, log-log paper (2 cycles). Note that there is a relatively high correlation (about 0.90) between the two measurements.

The calculation procedure is given for two cases: (1) assuming the data to be normally distributed (actually a bivariate normal) and (2) assuming the data are lognormally distributed. (In the latter case, only the changes in the calculations from case 1 are given.)[21]

Calculations (Assuming Bivariate Normal)

1. Calculate the mean and the standard deviation for each variable.

$\bar{x} = 56.5$ $\qquad$ $\bar{y} = 49.0$

$s_x = 26.3$ $\qquad$ $s_y = 22.9$

2. Calculate the correlation coefficient, r, for the two measurements,

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{s_x s_y} = 0.91 \qquad \text{Equation 3-19}$$

and n is the number of paired observations.

TABLE 3-11. TSP DATA FROM SITES 397140014H01 AND 397140020H01 FOR INTERSITE CORRELATION TEST, 1978

Site 20 (x), $\mu g/m^3$	Site 14 (y), $\mu g/m^3$	ln x	ln y
43	31	3.76	3.43
40	34	3.69	3.53
24	13	3.18	2.56
31	40	3.43	3.69
50	49	3.91	3.89
13	19	2.56	2.94
65	79	4.17	4.37
54	39	3.99	3.66
58	51	4.06	3.93
53	46	3.97	3.83
77	72	4.34	4.28
59	49	4.08	3.89
75	72	4.32	4.28
36	33	3.58	3.50
28	18	3.33	2.89
30	24	3.40	3.18
31	24	3.43	3.18
57	47	4.04	3.85
41	32	3.71	3.47
65	66	4.17	4.19
31	28	3.43	3.33
69	68	4.23	4.22
60	74	4.09	4.30
87	83	4.47	4.42
76	80	4.33	4.38
33	37	3.50	3.61
73	69	4.29	4.23
57	55	4.04	4.01
36	28	3.58	3.33
65	51	4.17	3.93
40	42	3.69	3.74
88	53	4.48	3.97
71	56	4.26	4.03
175	129	5.16	4.86
85	64	4.44	4.16

(continued)

TABLE 3-11 (continued)

Site 20 (x), $\mu g/m^3$	Site 14 (y), $\mu g/m^3$	ln x	ln y
56	46	4.03	3.83
88	46	4.48	3.83
75	57	4.32	4.04
33	26	3.50	3.26
50	41	3.91	3.71
57	41	4.04	3.71
108	90	4.68	4.50
46	31	3.83	3.43
69	63	4.23	4.14
42	37	3.74	3.61
76	69	4.33	4.23
101	108	4.62	4.68
38	37	3.64	3.61
57	47	4.04	3.85
45	43	3.81	3.76
28	26	3.33	3.26
33	25	3.50	3.22
52	45	3.95	3.81
38	39	3.64	3.66
43	23	3.76	3.14

3. Obtain a probability ellipse, as described in steps 4 through 7 and shown in Figure 3-9.

4. Assume that the means, standard deviations, and correlation coefficient are known values describing all data from the sites. There should be considerable representative data used in the determination of these statistics, at least 50 days (measurements) for each site. This is similar to the assumption made in the development of a quality control chart.

5. Assume a probability level such as 95%; that is, the ellipse to be constructed should contain 95% of the data points.

6. Determine a critical value of χ^2 (chi-square) for two degrees of freedom for 95% probability, see Table 3-12 for a list of values of χ^2 for various probability levels from 50% to 99.95%. The value is 5.99 for 95% probability.

7. Plot the ellipse with coordinates (x, y) which satisfy the equation,

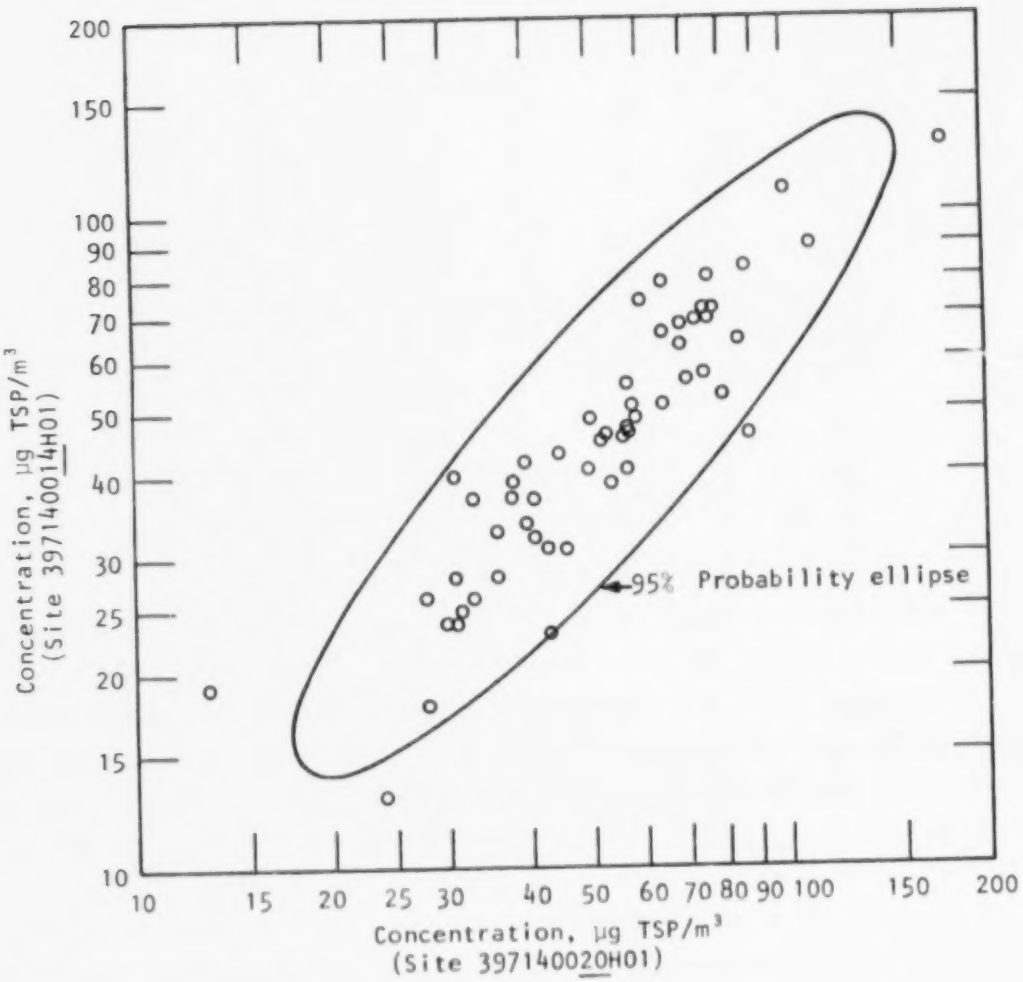


Figure 3-9. Intersite correlation test data.

$$\frac{1}{1-r^2}\left[\left(\frac{x-\bar{x}}{s_x}\right)^2 - 2r\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right) + \left(\frac{y-\bar{y}}{s_y}\right)^2\right] = \chi^2, \quad \text{Equation 3-20}$$

TABLE 3-12. χ^2 VALUES FOR TWO DEGREES OF FREEDOM, FOR VARIOUS PROBABILITY LEVELS

Confidence level	χ^2
50	1.39
60	7.83
70	2.41
80	3.22
90	4.61
95	5.99
97.5	7.38
99	9.21
99.5	10.6
99.9	13.8
99.95	15.2

or on substitution,

$$\frac{1}{1-(0.91)^2}\left[\left(\frac{x-56.5}{26.3}\right)^2 - 2(0.91)\left(\frac{x-56.5}{26.3}\right)\left(\frac{y-49.0}{22.9}\right) + \left(\frac{y-49.0}{22.9}\right)^2\right] = 5.99.$$

This ellipse falls within (inscribed within) a rectangle with center at $(\bar{x}, \bar{y})$ and with two sides of lengths $2s_x \chi = 2s_x \sqrt{5.99}$ and $2s_y \chi = 2s_y \sqrt{5.99}$, respectively. In this example,

$$2s_x \sqrt{5.99} = 128.9$$

$$2s_y \sqrt{5.99} = 112.0.$$

Note that the $\sqrt{5.99}$ is the square root of the χ^2 value corresponding to the confidence level selected.

The computations are tedious and should be programmed for computerized solution on repeated use. Another approach is to use the ellipse plotting procedure adapted for the NORMIX cluster analysis.[14]

Calculations (Assuming Lognormal Distribution)

If the analysis is performed using the logarithms of x and y (say x′ and y′) one can substitute x′ and y′ throughout the above steps. After all of the computations have been completed using the logarithms, transform the results back to the original

data by using either $e^{x'}$ or $10^{x'}$, depending on whether natural or common logarithms have been used.

The computations for the logarithms are:

$$\bar{x}' = 3.94 \qquad \bar{y}' = 3.79$$

$$s_{x'} = 0.44 \qquad s_{y'} = 0.47$$

$$r = 0.90$$

The lengths of the sides of the rectangle are:

$$2s_{x'} \cdot \sqrt{5.99} = 2.15$$

$$2s_{y'} \cdot \sqrt{5.99} = 2.30.$$

If a data point falls outside of the ellipse then the values for both x and y should be flagged for checking.

There are four data points outside the 95% confidence ellipse. If the data had been studied one variable at a time, the point (175,129) would be subject to question, as indicated in Section 3.3.4. Even though this point falls outside the ellipse it does not appear to be inconsistent with the remaining data based on data from both of the sites; however, further study of these relatively high values would be advisable. The other three data points have at least one coordinate value within the range of the other data, and studying one variable at a time would not necessarily suggest these coordinate values as possible anomalies. Two of these points should be studied further, based on the correlated data.

3.5 REFERENCES

1. U.S. Department of Commerce. Computer Science and Technology: Performance Assurance and Data Integrity Practices. National Bureau of Standards, Washington, D.C. January 1978.

2. Data Validation Program for SAROAD, Northrup Services, EST-TN-78-09, December 1978, (also see Program Documentation Manual, EMSL).

3. Barnett, V., and T. Lewis. Outliers in Statistical Data. John Wiley and Sons, New York, 1978.

4. U.S. Environmental Protection Agency. Screening Procedures for Ambient Air Quality Data. EPA-450/2-78-037, July 1978.

5. W. F. Hunt, Jr., T. C. Curran, N. H. Frank, and R. B. Faoro, "Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air," Transactions of the Joint European Organization for Quality Control/ International Academy for Quality Conference, Vernice Lido, Italy, September 1975.

6. W. F. Hunt, Jr., R. B. Faoro, T. C. Curran, and W. M. Cox, "The Application of Quality Control Procedures to the Ambient Air Pollution Problem in the USA," Transactions of the European Organization for Quality Control, Copenhagen, Denmark, June 1976.

7. W. F. Hunt, Jr., R. B. Faoro, and S. K. Goranson, "A Comparison of the Dixon Ratio Test and Shewhart Control Test Applied to the National Aerometric Data Bank," Transactions of the American Society for Quality Control, Toronto, Canada, June 1976.

8. Grubbs, F. E., and G. Beck. Extension of Sample Sizes and Percentage Points for Significance Test of Outlying Observations. Technometrics, Vol. 14, No. 4, November 1972.

9. Curran, T. C., W. F. Hunt, Jr., and R. B. Faoro. Quality Control for Hourly Air Pollution Data. Presented at the 31st Annual Technical Conference of the American Society for Quality Control, Philadelphia, May 16-18, 1977.

10. Johnson, T. A Comparison of the Two-Parameter Weibull and Lognormal Distributions Fitted to Ambient Ozone Data. PEDCo Environmental, Inc., Durham, North Carolina, and The Air Pollution Control Association. Quality Assurance in Air Pollution Measurement. Presented at the Air Pollution Control Association, New Orleans, March 11-14, 1979.

11. Larsen, R. I. A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards, Publication No. AP-89, U.S. Environmental Protection Agency, 1971.

12. Marriott, F. H. C. The Interpretation of Multiple Observations. Academic Press, New York, 1974.

13. Hawkins, D. M. The Detection of Errors in Multivariate Data Using Principal Components. Journal of the American Statistical Association, Vol. 69, No. 346. 1974.

14. Wolfe, J. H. NORMIX: Computation Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions, Research Memorandum SRM 68-2. U.S. Naval Personnel Research Activity, San Diego, 1967.

15. U.S. Environmental Protection Agency. Guidelines for Air Quality Maintenance Planning and Analysis. Vol. 11. Air Quality Monitoring and Data Analysis. EPA-450/4-74-012, 1974.

16. U.S. Environmental Protection Agency. Quality Assurance and Data Validation for the Regional Air Monitoring System of the St. Louis Regional Air Pollution Study. EPA-600/4-76-016, 1976.

17. W. F. Hunt, Jr., J. B. Clark, and S. K. Goranson, "The Shewhart Control Chart Test: A Recommended Procedure for Screening 24-Hour Air Pollution Measurements," J. Air Poll. Control Assoc. 28:508 (1979).

18. Grant, E. L., and R. S. Leavenworth. Statistical Quality Control. McGraw-Hill Book Company, New York.

19. Siegel, S. Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, 1956.

20. Remington, R. D., and M. A. Schork. Statistics with Applications to the Biological and Health Sciences. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970.

21. Hald, A. Statistical Theory with Engineering Applications. New York, 1952.

4.0 SELECTION AND IMPLEMENTATION OF PROCEDURES

There are several factors to be evaluated before one can select the most appropriate data validation procedure for the specific application. These factors can be categorized into two sets of decision criteria (1) those based on an organizational perspective and (2) those based on analytical characteristics. Table 4-1 gives a breakdown of the factors to be considered in the selection of data validation procedures.

TABLE 4-1. FACTORS TO CONSIDER IN THE SELECTION OF DATA VALIDATION PROCEDURES

Organizational Criteria
1. Number of data sets
2. Historical data requirements
3. Nature of data anomalies
4. Manual methods
5. Continuous methods
6. Strip chart data
7. Magnetic tape data
8. Data transmitted by telephone lines
9. Timeliness of the procedure
Analytical Criteria
1. Statistical sophistication
2. Computational requirements
3. Expense of analysis
4. Sensitivity of the procedure
5. Use of data

The organization criteria can be used as an initial screening of the procedures based on the data needs, the analytical capabilities of the agency, and staff training. For example, a local agency with limited staff and without computer facilities and statistical support would select from those procedures in Sections 3.1.1, 3.1.2, 3.1.3, 3.2.1, 3.3.1, and 3.3.2--that is,

data ID checks, unusual event review, deterministic relationship checks, data plots, gross limit checks, and pattern checks. On the other hand, a Federal agency with extensive capabilities can use any of the validation procedures with heavy emphasis on computerized procedures, computer graphics, Shewhart control charts, the gap test, and/or the Johnson p test.

The factors are described in Sections 4.1 and 4.2. Table 4-2 gives a selection procedure using three scenarios based on whether or not a local, State, or Federal agency has computer and statistical resources available. Section 4.3 contains a discussion of the implementation of the data validation process.

TABLE 4-2. SELECTION OF DATA VALIDATION PROCEDURES

Scenario of Agency/Resources	Select procedures from those in Subsections listed below
State or local agency	
1. Without both computer and statistical support	3.1.1, 3.1.2, 3.1.3, 3.2.1 (limited to manual effort) 3.3.1, 3.3.2
2. With computer but limited statistical support	The above plus 3.2.1 (computerized graphics), 3.2.2 3.3.3, 3.3.4, 3.4.1
Federal, State, or local agency With both computer and statistical support	Any procedure in Section 3.

4.1 ORGANIZATIONAL CRITERIA

The selection criteria which can be used in preliminary screening of the validation procedures are given in this section.

4.1.1 Number of Data Sets

The test procedures for internal consistency are designed for the validation of a single data set, without the use of historical data. If there is an instrument error which consistently alters all of the values in the data set, these test procedures will not be of any value. Hence it is necessary to use at least

one test which identifies or flags data based on previous or historical data or on comparison with other (parallel) data sets. If an extensive analysis of past data is available or can be readily performed then the test procedures of Section 3.3 should be considered, otherwise the data should be compared with other parallel data sets which are described in Section 3.4.

4.1.2 Historical Data Requirements

The major test procedures which require historical data for setting limits are the gross limit checks and pattern tests, the Shewhart control chart, and the intersite correlation test. In all of these, data are required over an extensive time period so that the results can be applied to new data. Limits can be altered as appropriate, to take into account both the additional data and the experience gained in using the procedures.

4.1.3 Nature of Data Anomalies

Some tests are designed for a single outlier (extreme value), whereas other tests are designed to detect shifts in either the variation of the data and/or in the mean (or median). For example, there are Dixon ratio tests for a single outlier and an outlier pair; the sign test checks for shifts in the mean or median. The Shewhart test detects shifts in either the mean or the variance (range or standard deviation) and anomalous outliers by means of the range check. The graphical techniques are likely to identify any one of the data anomaly types if the graph is not too complex.

4.1.4 Strip Chart Data

The analysis of strip chart data can be limited to simple visual scans for gross anomalies. However, it may also be desirable to establish limits within which successive values must fall for the data to be valid. Hence, the successive difference test would be particularly useful. Also, gross limit and pattern tests can be easily applied to strip chart data.

4.1.5 Size of Data Set (Manual and Continuous Methods, Table 4-1)

Manual methods are most appropriate for small data sets, typically daily averages for each day or for every sixth day. Procedures which can be performed without a computer are the Dixon ratio, the Shewhart control chart, data plots, and the routine procedures described in Sections 3.1.1, 3.1.2, and 3.1.3.

Computer methods are desirable for validating large data sets (e.g., from continuous monitoring of hourly average concentrations) since data handling and calculations by manual means can become tedious and can result in errors. However, the data validation procedure can be added to the standard analysis and data storage procedures to minimize the total computer time.

4.1.6 Magnetic Tape Data

Data recorded on magnetic tape can be easily validated using computerized methods. Almost all of the procedures can be considered. The more sensitive and economical ones for the expected types of anomalies would be best used. Turnaround time should be minimal. Section 3.1.4 describes several procedures which should be routinely used to maintain data integrity.[1] For example, the data can be blocked into logical periods (days, weeks, months, etc.) and test performed for data sets using internal consistency checks (e.g., Dixon ratio) or comparison with historical data (e.g., control chart). In addition routine checks can be made of the descriptive identification codes.

4.1.7 Data Transmitted by Telephone Lines

The recommendations for data transmitted by telephone lines are the same as those for data on magnetic tape. Experience gained through CHAMP suggests it is not always advisable to use the telemetry data when the magnetic tape data are available and when the latter can be used in a timely manner. In any case, there should be data validation checks to ensure that the telemetry data are the same as the data on magnetic tape and/or the raw data. Section 3.1.4 describes several procedures to be considered for routine use.[1]

The principal result of transmission error is the loss or alteration of data. One way to check for transmission error is to transmit the data a second time and then pair the two data streams. Gaps and alterations will be immediately apparent unless the transmission error is systematic.

4.1.8 Timeliness of Procedure

One of the key aspects of a good data validation program is the timely identification of data anomalies and the feedback of this information to the data source for corrective action and for review of the raw data and background information needed to check the validity of the suspect data. The shortest possible turnaround time is desired so that the information loop can be closed before vital information is lost and/or before more questionable data are generated. The combination of procedures selected must satisfy these time constraints. For example, some procedures with very quick response times can be used jointly with more complex procedures in order to catch gross errors quickly and to catch less obvious errors within a slower response time but still within a time frame that can result in useful data validation implementation.

4.2 ANALYTICAL CRITERIA

The selection factors which are analytical in nature are described in this section.

4.2.1 Statistical Sophistication

The test procedures are arranged within each subsection of Section 3 from those that are least sophisticated to those that are most complex. Although some test procedures may appear to be statistically complex, these same test procedures may be very convenient to use. On the other hand, some of those with least sophistication can require considerable effort (e.g., plotting of all of the data).

4.2.2 Computational Requirements

Although a computerized approach is desirable for several of the procedures, all of them can be accomplished by manual means.

The only procedures requiring computer help for routine use with large volumes of data, are data plots, gap and Johnson p tests, intersite correlation, and rank sum tests. For small data sets, many of these tests can be performed easily by manual means. Some procedures require considerable computation to derive the limits from historical data, but the test is easily performed after this initial step, which need not be repeated until one suspects that the limits need to be changed to reflect real changes in air quality.

4.2.3 Expense of Analysis

The expense of the analysis in manual/computer time is relatively small compared to data costs and/or the cost of providing invalid data to a data bank for use by a decision maker. However, these costs must be considered from the standpoint of using an efficient data validation procedure. If a procedure ignores too many data anomalies or flags too many good data, either the limits need to be adjusted or the procedure eliminated from further use. The expensive procedures are not always the best (e.g., the Shewhart test will not cost much on a continuing basis, and it will detect many types of invalid data). The graphical plots can be very time consuming, by manual means or by computer; however, the results will also prove very useful for detecting invalid data that would be missed by other procedures.

4.2.4 Sensitivity of the Procedure

The capability of the data validation procedure to correctly identify invalid data, is critical in the selection. One should also refer to Section 4.1.3 above which discusses the nature of the data anomaly. Some procedures will identify a single outlier; others will detect shifts in the median (or mean) level, while still others will detect changes in the variability of the data.

Some procedures, such as the Dixon ratio test for a single large value (Section 3.2.2.1), may be very insensitive if there is a second large value. In these cases, another Dixon ratio

test (Section 3.2.2.2) can be used which is sensitive to pairs of outliers. It is desirable that the user of the procedure determine the types of invalid data and then design the data validation program to discriminate for these specific types of data. Thus it is not possible to say that one procedure is always best, rather, the appropriateness of a procedure varies with the type of data anomaly.

Decisions concerning sensitivity requirements can be made best after some experience has been gained in using the procedure. The data analyst should continually note the percentage of flagged data which upon investigation are found to be erroneous values. A control chart may be used to maintain a record of the performance of the data validation procedure. The sensitivity and cost of implementation of different procedures should be compared. In comparing the sensitivities one must be careful to obtain an independent assessment of each procedure. This may be a problem when the procedures are applied successively (or in series). An exponential identifier (e.g., see Section 5.5.1 for the use of 10^{34} for invalid data) can be used in computer applications to indicate the procedure(s) which identified the data as questionable and to aid in statistical analysis of the efficiency of the procedures.

4.2.5 Use of Data

The intended use of data should be considered in determining the time and level of effort to be put into data validation. If the data are to be used in making important policy decisions, then they should be more carefully screened than if the data are being used in a less critical role. If in doubt, err on the side of more careful validation; data are often used in a manner unknown to the generator/originator of the data.

4.3 IMPLEMENTATION OF DATA VALIDATION PROCESS

To make the data validation process effective requires the consideration of several important factors in both the selection and implementation of the procedures. Some of the factors to be

considered in implementation were identified in Section 2.4 and are discussed in this section.

Initially, a person needs to be identified with the responsibility for the data validation activities. In a local or State agency, this may be a part-time function; whereas in a regional or Federal agency, one or more persons may be given this responsibility on a full-time basis.

The data validation procedures need to be selected by taking into consideration the several factors identified in Sections 4.1 and 4.2. The process must be consistent with, for example, the use of the data, the network size, data volume, and staff size and training.

After the selection of the procedures, the limits, patterns, and confidence levels must be determined (using historical data) for a particular application to obtain an efficient procedure. The limits may need to be adjusted after some experience with the procedures. The tighter the limits, the greater the quantity of data that are flagged and hence the greater is the effort of checking on the flagged data. There is a tradeoff between the stringency of the data validation and the costs of flagging too many data values (and the completeness of the remaining data, particularly if the data identified by the validation process are deleted or flagged relative to further use in some computations).

A data validation plan should be prepared and documented as a part of the quality assurance plan (as described in Volume I of the Quality Assurance Handbook).[2] This plan should include:

1. The data flow (reporting) hierarchy;
2. The data validation check points;
3. The methods for checking the data;
4. The documentation of the process (the data values flagged for investigation, the values inferred to be anomalies and those for which no clear decision could be made);
5. The techniques used for anomaly investigation;
6. The reporting schedules;
7. The mechanism for feedback of validation information to the data collection, to the quality control procedures, and to the validation process;

8. The resources (including costs) for performing data validation;[3]

9. Listing of QC checks used to reject data forthrightly and without further investigation (e.g., the use of zero and span checks to reject data when the span drift exceeds a specified limit);

10. The routine handling of data in blocks (or time periods) with a corresponding sign-off form indicating the data have been subjected to the data validation process; and

11. The procedures for tracking the causes of invalid data.

A periodic reporting procedure must be used to summarize the performance of the data validation process with regard to this plan. Finally, the validation process can be made most effective by adjusting, if necessary, the limits (or criteria) in order to minimize the incorrect number or percentage of inferences with regard to possible data anomalies. Each investigation costs time and money, so the incorrect inferences must be minimized. A good documentation system should enable the validator to solve this problem.[3] An effective reporting and feedback system will also contribute to the motivation of the personnel involved with the process and to timely corrective action.

4.4 REFERENCES

1. U.S. Department of Commerce. Computer Science and Technology: Performance Assurance and Data Integrity Practices. National Bureau of Standards, Washington, D.C., January 1978.

2. U.S. Environmental Protection Agency. Quality Assurance Handbook. Vol. I. Principles. EPA-600/9-76-005, 1976.

3. Smith, F. Guideline for the Development and Implementation of a Quality Cost System for Air Pollution Measurement Programs. Research Triangle Institute. RTI/1507/01F, November 1979.

5.0 HYPOTHETICAL EXAMPLES AND CASE STUDIES

Two hypothetical examples and three case studies are described in this section. The two examples are very simple ones involving ambient air monitoring and source testing and are given first in Sections 5.1 and 5.2. The three case studies are given last and in increasing order of sophistication (Sections 5.3, 5.4, and 5.5). Few simple examples as described in Sections 5.1 and 5.2 are documented in quality assurance or pollution control literature because the individuals involved probably do not feel their programs reveal any innovative procedures. Hence these examples are hypothetical, though hopefully realistic of the needs of the local agency performing ambient air monitoring and source tests.

5.1 HYPOTHETICAL EXAMPLE FOR AMBIENT AIR MONITORING

Suppose that a local agency has four hi-vol monitoring stations for TSP, operating every sixth day, and two stations for ozone, operating 5 months per year. At one station, temperature, rainfall, and wind data are being obtained. Two cases will be considered: one without and one with computerized methods. In each case the preferred procedures will be selected for the data validation process.

5.1.1 Without Computerized Support - Several very simple and useful manual procedures which can be used to validate the data will be suggested for each pollutant/parameter.

TSP - There are approximately 60 measurements per year at each site. If all four sites have different monitoring schedules, the data from each site should be validated independently. A gross limit check should be established for TSP, perhaps the air quality standard value if this value has a small probability of occurrence, say less than 5%. A gross limit check value can be obtained by analyzing the historical data for the most recent

year and determining a single value which has a small likelihood of exceedance. Gross limit checks are discussed in Section 3.3.1. In some cases, multiple limits or a pattern test should be used depending on the results of the following analysis for each of the four stations

1. Analyze 1 year of data (about 60 values), examine the data for a seasonal/day-of-week pattern (e.g., four seasons, weekday and weekend day, eight categories),

2. If there are no significant patterns, then use a single gross limit check; otherwise, determine a limit for the indicated pattern, say weekend versus weekday values.

Typical TSP data can be adequately approximated by the lognormal distribution, hence it is usually necessary to make the lognormal transformation of the data and then to statistically analyze the transformed data (including the analysis described above) and ultimately to transform back (if desired) to the actual data values to obtain the appropriate limits. The limits determined in this way are usually larger but more statistically accurate than if no transformation were used.

A second procedure, which is usually beneficial in detecting changes in the data from one or more of the four hi-vol stations, is to develop a quality control chart for the following:

1. The hi-vol, average and range values for each set of five consecutive readings (about 1 month of data) for each of the four stations, and

2. The average and range for the four station averages.

The four charts developed in step 1 monitor the individual stations for suspicious changes in either the average or range. The last chart monitors the group of four stations for suspicious changes in the overall average and for the variation among the four station averages. To develop the control limits for these charts, use the data for 1 year (60 values, or 12 sets of five each) and the methodology described in Appendix H, Volume I of the Quality Assurance Handbook and in Appendix C of this report.[1]

Finally, a year of data can be plotted in a few hours since there are only 240 measurements per year (60 values at four sites). The data for all four hi-vols should be plotted on the same graph or on graphs placed one below the other to facilitate comparisons.

O_3 - The two O_3 sites are assumed to be in operation for 5 months of the year and to record hourly average values. For these data, the preferred data validation procedures would be gross limit and pattern checks of Section 3.1. Limits such as those in Table 3-2 of Section 3.3 should be derived from historical data for at least 3 years if available. The data should be statistically analyzed for day-night patterns and perhaps for two groupings within the 5 months--that is, the 3 months with higher O_3 levels and 2 months with lower O_3 levels. Each of the 5 months also could be treated separately in the analysis and then grouped as the data and the statistical analysis indicate.

The Shewhart quality control chart may be used to monitor both stations. The historical data would need to be grouped in the same manner as they are to be used in the quality control chart and then analyzed for averages and ranges to obtain the quality control limits. Note: The data do not have to be plotted on charts to apply this procedure; however, the visible chart may be beneficial in detecting trends/patterns in the data that would not be evident on review of a data record.

Meteorological Data - Assume that hourly average temperature, rainfall, windspeed, and wind direction are available for the data validation.

The following pattern tests should be used with necessary modifications to be consistent with the meteorological conditions for the agency:

1. The minimum hourly average should occur between 04-09 hours, and the maximum between 12-17 hours.

2. Hourly changes should not exceed 6°C (10°F). If a decrease of 6°C (10°F) or more occurs, the wind direction should have changed to a more northerly direction, and/or rainfall of 0.15 in. or more should have occurred.

3. Hourly values should not fall outside of the specified maximum and minimum values for each month. These limit values need to be determined from historical data.

5.1.2 With Computerized Support -

TSP - The routine manual procedures suggested in Section 5.1.1 may be computerized if desired. No additional procedures are recommended because the data are not voluminous and because the recommended procedures can be performed manually.

O_3 - In the case of O_3 the volume of the data makes the use of a computer worthwhile. For example, a plot of the O_3 data by hour can reveal any unusual values and/or changes in consecutive values from the typical diurnal and seasonal patterns.

In addition, a check of the extreme values for 5 months using the gap or "Johnson" p test (Sections 3.2.4 and 3.2.5) could be beneficial. Particularly in the Johnson p test, the distribution of the upper portion of the data is approximated in the process of checking the extreme values. Hence the identification of a possible data anomaly may actually be of secondary interest in the use of the Johnson p test. The computations can be easily performed on a programmable calculator or a computer given the data values and procedure(s) in Section 3.2.5 and Appendix B.

Meteorological Data - The availability of a computer should not materially affect the selection of the procedure; however, it may result in the use of a computerized method of performing the routine manual analyses.

5.2 HYPOTHETICAL EXAMPLE FOR SOURCE TESTING

Suppose that a Method 6 source test is performed at a utility plant with a flue gas desulfurization (FGD) system. One of the obvious data validation checks to perform is a report review using gross limit and/or parameter relationship checks. Some data checks to be performed are:

1. Barometric pressure - At sea level, the value should be approximately 760 mm (30 in.) and between 736 and 787 mm (29 and 31 in.) Hg; at other elevations, the value should decrease 2.8 cm (1.1 in.) Hg per 305 m (1000 ft) above sea level.

2. Moisture data - Nomographs provide moisture content at saturation as a function of stack absolute pressure and stack gas temperature. If the reported value is higher than the maximum that was read from the nomograph, the data are suspect.

3. Volumetric flow rate data - These data are difficult to cross-check accurately, but gross limits can be determined from engineering considerations.[2] In addition, the data should be compared with data obtained under similar process conditions from previous tests on the same source, similar sources, or tests performed at the inlet to the control device.

4. Emission results - These are the most difficult but also the most important data to validate. A sulfur balance should yield a good check of the SO_2 emission results. However, there is considerable variability in the sulfur in coal (relative standard deviation of 15% is not unusual). The limited amount of data for a series of three runs does not make the use of a Dixon ratic or Grubbs test practical.[2] That is, a decision may be made to eliminate one extreme from a group of three values (using the Dixon test) as a result of process variations of the order of two to one resulting from factors beyond the control of the plant. Hence, decisions concerning the validity of data need also be based on other experiences with this or similar plants.

5.3 CASE STUDY OF A MANUAL DATA VALIDATION SYSTEM

This section describes an automated monitoring and data acquisition system for which the validation procedures are performed manually. The monitoring system includes 26 sites which monitor 124 environmental parameters. In the case of one parameter, SO_2, a comprehensive formal quality assurance program has been developed. In the case of the other parameters, no formal quality assurance procedures have been documented. A brief description of the validation procedures developed for the parameters without the quality assurance programs will be followed by a more detailed description of the validation procedure for SO_2. The case study illustrates the complementary nature of

the quality control and the data validation procedures within a monitoring system, and it shows clearly that, as more comprehensive quality control procedures are applied during data generation, the need for in-depth data validation procedures can be limited to less frequent review or to spot check items listed under QC below. The network to which this case refers is not identified for proprietary reasons.

5.3.1 Quality Control

The major QC procedures for the instruments other than the continuous SO_2 monitors are: daily zero-span checks, 6-month multipoint calibration, and 6-month scheduled preventive maintenance. Instruments are rotated in the field. Each instrument is brought to the central laboratory for maintenance and calibration every 6 months and is replaced in the field by another instrument (a better practice would be to calibrate the instruments in the field). Instruments are checked on site by technicians two or three times a week on the average. Data are transmitted automatically to a central computer. Strip charts are generated as backup to the computer data base.

A total quality assurance program was designed for the SO_2 monitors in this network. Field level QC procedures include the following:

1. Operator checklists - Upon entering a monitoring shelter, the field technician completes a checklist of key parameter observations: pressure gauge readings, critical temperature readings, weather or site-specific conditions, and so forth, that may significantly influence SO_2 data quality. Checklists are designed to instruct the operator to take specific action under specific parameter conditions--for example, if a parameter is out of tolerance; action may include phone consultation with the chief technician. Checklists are forwarded to the laboratory weekly.

2. Zero-span control charts - The field operator checks the SO_2 strip charts and manually plots the zero-span results for each day on a control chart. Predetermined control limits are

established for each zero-span control chart. Control charts are forwarded to the central laboratory for review and approval monthly. If an out-of-limits condition occurs during the month, the field technician consults by phone with the chief technician.

3. Maintenance reports - Formal reports of unscheduled maintenance performed in the field are completed as required and forwarded weekly to the chief technician for review. This procedure facilitates identification of chronic instrument problems.

4. Strip chart review - The field operator looks for unusual traces on the strip chart that indicate equipment malfunctions. If malfunctions are suspected, the operator makes a note on the strip chart and flags the problem for the chief technician. Strip charts are forwarded to the central laboratory for review approximately monthly.

5. Calibration and preventive maintenance - Instruments are calibrated and maintained on a rotating 6-month schedule. (These schedules were selected prior to recent quality assurance standards which recommend a quarterly schedule.)

6. Laboratory strip chart scan - The quality assurance coordinator scans SO_2 strip charts upon receipt from the field to identify uncharacteristic traces that may have been missed by the field technician. For time periods when either the field operator, the chief technician, or the quality assurance coordinator know that the data are unacceptable, a form is completed to inform a data validator to delete the data from the data base. A data delete form is only prepared when there are backup data or information (e.g., zero/span data, remark in the daily log; etc.) to justify the action.

7. Quality assurance filing system - All maintenance and calibration records, zero-span control charts, field checklists, strip charts, and data delete or change request forms are marked with the site and instrument identification. These documents are filed permanently by site, instrument, and date to provide future backup for answering data validation questions.

In addition to the QC procedures above, the central air quality monitoring laboratory has a teletype capability to poll any monitoring site for specific parameters.

8. Data review and audit - The quality assurance coordinator performs an annual audit of the files and traces the occurrence of events from causes through subsequent corrective actions. Also the laboratory supervisor reviews the data delete forms on a monthly basis to suggest corrective actions.

5.3.2 Data Validation

The actual data validation in this system is performed by a contractor not associated with the routine operation of the system. The validation contractor's responsibilities include:

1. Strip chart review,
2. Interparameter relationship review,
3. Recapture of valid data lost during data transmission,
4. Confirmation of excursions above the ambient air quality standards, and
5. Deletion from the data file of all data invalidated through the QC or validation procedures.

The following two exceptions apply:

1. Strip chart review is not performed since the QC procedures for SO_2 specify that this technique be applied at the field level. Although this is the only real point of departure between validation of SO_2 data and validation of the other parameters, it is significant because an in-depth, after-the-fact review of strip charts is time consuming and expensive for a network of this size. For example, after-the-fact chart review for SO_2 would consume on the average of 1 to 2 hours of technician time per site-parameter-month combination, depending on the kinds of problems encountered and the volume of communication required between the data validator and the monitoring laboratory.

2. Interparameter relationship review is applied to meteorological data only, except in the case of excursions above the ambient air quality standards.

5.3.2.1 Strip Chart Review - Strip charts are scanned manually to identify questionable data periods. The purpose of the scan is to identify:

1. Unusual spikes,
2. Invalid data periods identified by field technicians, and
3. Uncharacteristic traces, for example a relatively flat or constant windspeed trace over several hours.

5.3.2.2 Interparameter Relationship Review - As indicated previously, the interparameter relationship review is performed for meteorological parameters only. Hourly averages are reviewed from formatted computer reports that display data for each hour of each month. One review compares dewpoint and ambient temperatures and then flags hourly dewpoint values that exceed the hourly temperature. Daily average dewpoint values that exceed the daily average temperature are also indicated. Another review considers windspeeds measured at two tower levels. Data are flagged whenever the average windspeed at the lower level exceeds that at the upper level. Procedures like these must be based on judgments by an experienced meteorologist. Validation instructions in many cases direct the validator to backup strip charts for questionable time periods identified during the interparameter relationship review.

5.3.2.3 Recapture of Valid Data Lost During Data Transmission - The data validation contractor receives a magnetic tape of data collected through the data transmission system at the same time the strip charts, data change requests, and data listings needed for validation are received. The objective of the monitoring system is to reach a level of at least 90 percent valid data collection. The data validator reviews the data listings for time periods when data capture is below 90 percent. The validator then checks the strip charts for those time periods and reduces data for specific hours that appear to be valid but were not picked up by the data transmission system. The data values are picked up from the strip charts through an electronic data digitizer and then used to update the data file on the magnetic tape.

5.3.2.4 Confirmation of Excursions above the Ambient Air Quality Standards -The data validator receives a listing showing specific hour, 3-hour, or 24-hour values that constitute excursions above the ambient air quality standards. The strip charts are examined for these specific time periods to verify the reported values. If the reported data differ from the strip chart values and the strip chart values are valid, then the data are changed in the file. The standards report is annotated and returned to the client's quality assurance coordinator for filing.

The confirmation procedure requires the data validators to check meteorological data records and strip charts for the time periods when standards excursions were reported. Unusual conditions or invalid meteorological data are also recorded on the standards report.

5.3.2.5 Deletion of Invalid Data - The final step in the validation procedure is the update of the data tape. Data picked up during data validation, as explained in Section 5.3.2.3, are merged into the file. During the update procedure, data found to be invalid by either the data validators, by the field operators through the QC procedures, or by the quality assurance coordinator (in the case of SO_2) are deleted from the file. A listing of each value added to or deleted from the permanent file is produced for filing with each monitoring stations records.

5.4 CASE STUDY OF THE CHAMP AUTOMATED DATA VALIDATION SYSTEM

The U.S. EPA and the Rockwell International Science Center designed and implemented an automated data acquisition system for the Community Health Air Monitoring Program (CHAMP). CHAMP was implemented to provide reliable air quality data to support the Community Health Environmental Surveillance System (CHESS), a program to study the effects of air pollutants on community health.

The CHAMP network included 23 monitoring sites in six cities. The network provided data for the 19 environmental parameters listed in Table 5-1. Monitors at each site were polled by minicomputers at the sites, and the central computer at

TABLE 5-1. CHAMP ENVIRONMENTAL PARAMETERS

Computer code	Parameter
NO_x	nitrogen oxides
NO	nitric oxide
NO_2	nitrogen dioxide (calculated)
SNO	sampled nitrogen oxide
O_3	ozone
SO_2	sulfur dioxide
CH_4	methane
THC	total hydrocarbons
NMHC	nonmethane hydrocarbons
WS	wind speed
VWM	vector wind magnitude
VWD	vector wind direction
TOUT	temperature outside
TIN	temperature inside
DEWP	dewpoint
BP	barometric pressure
HVFL	hi-vol flow
RSPI	respirable suspended particulate flow rate; inlet
RSPF	respirable suspended particulate flow rate, final

Research Triangle Park (RTP), North Carolina polled the sites every 2 hours. These data were used to review system status at each remote monitoring site and to provide feedback for field personnel to perform maintenance as needed.

Magnetic tapes were generated at each site, and the tapes were forwarded to RTP every 2 weeks. The data validation procedures were applied to the data on tape rather than to the data transmitted via telephone lines to the central facility because the data on the tapes were generally more complete and more reliable than the telemetry data.

Strip charts for gaseous pollutants were generated at remote sites. The strip charts provided a backup, and data that were lost due to computer failure were recaptured from the strip charts. The strip charts also provided a visual QC check for station operators.

5.4.1 Quality Control Functions

Several automated QC checks were performed by the on-site computers as the data were generated. Manual QC procedures were used also.

The automated QC procedures were based on limits set for the analog values of the critical operating parameters. Critical limits were entered into the computer; if these limits were exceeded, then the associated 5-minute averages were flagged as possibly invalid. Direct environmental measurements, referred to as "primary channels," were checked for nonzero voltage, normal setting of all valves, power to the instrument, and digitally measured tolerances for proper ambient sampling. Instrument operations (e.g., hydrogen flow in an SO_2 monitor), referred to as "secondary channels," were also tested to verify that values were within tolerances; and status conditions necessary for correct sampling were machine checked every five minutes. The same control procedure was applied to the calibration operation to verify proper flows of the calibration gases and correct valve settings on the flow system. Figure 5-1 illustrates the QC procedure.

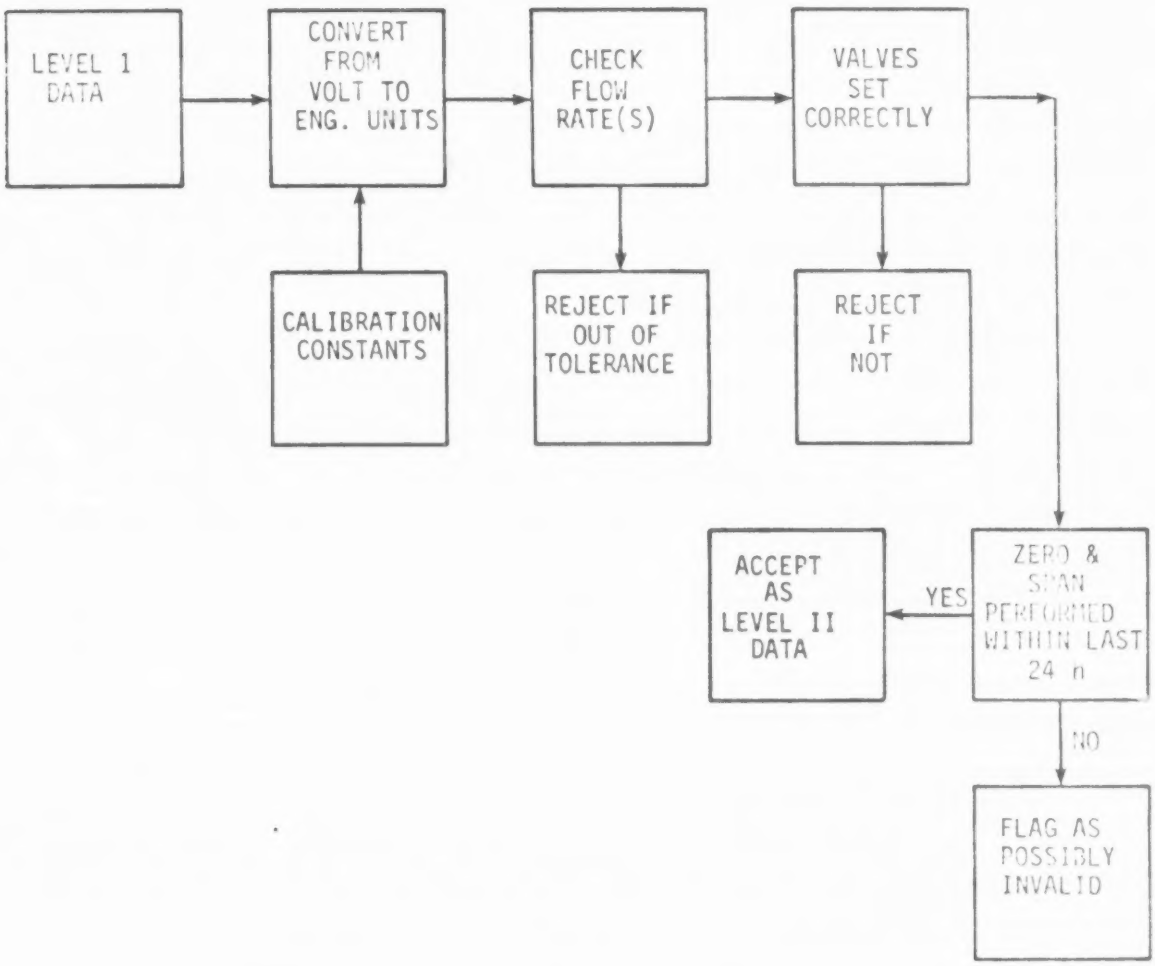


Figure 5-1. Automated quality control tests.

During calibration and instrument tests, the sampling indicator was in the off position. During sampling, the indicator was in the on position. During QC checking, the status bits were read to determine the on or off condition.

Error conditions flagged by the automated system were reviewed daily at the central computer facility. As a result of the QC review of data received, either personnel could be dispatched to remote sites for instrument adjustments or field operators could be advised of needed changes. This real-time turnaround afforded by automated QC checks was important for maintaining continuous operation throughout the system. It allowed needed adjustments to be made quickly so that a minimum quantity of data was lost or invalidated.

Manual QC checks included visual inspection of control charts for unusual values or large changes in reported data values. If the field operator noted unusual events at a site which might affect subsequent data validation or review, he or she was instructed to enter explanatory entries into the data file.

Manual QC calibration audits were performed on the entire system quarterly. Zero-span calibration was performed every 3 days. A tickler file was maintained at each site to indicate when preventive maintenance was required.

5.4.2 Data Validation

The data validation procedures were performed at the CHAMP central computer facility at RTP. Computer listings of the data that had been subjected to QC checks were reviewed.

A report listing the hourly average values for each of the 19 parameters (primary channels) measured at the remote monitoring stations was inspected. For each station, the time (hour), the parameter name, the parameter hourly average value, and the number of valid 5-minute values used to compute the hourly average were listed as shown in Figure 5-2. The data shown for ozone for hours 13 through 15 illustrate the application of this technique. For hour 13, twelve valid 5-minute values were used

to compute the hourly value of 0.1356. For hour 14, only four valid 5-minute values were used in calculating the hourly average of 0.1532. The "B" indicates that data were missing and invalid for the remaining eight 5-minute periods during the hour. Invalid data would be the result of primary or secondary channels being outside acceptable limits during the QC tests on site. During hour 15, only one valid 5-minute average was available. The "M" indicates that the remaining eleven 5-minute averages were missing.

As indicated in Figure 5-3, a summary of secondary parameters associated with each of the primary parameters in the previous listing was produced. The summary shows the high and low critical limits for each parameter. Note that in the case of ozone, the critical value range for the flow of ethylene (FETH) was set at 20.000 to 30.000.

A third listing, Figure 5-4, shows the reasons for invalid data flagged in the hourly data listing illustrated in Figure 5-2. Again using the ozone parameter, the listing indicates that FETH was outside of the critical limits during hour 14. Figure 5-5 shows the invalid 5-minute values for ozone and the 5-minute values of each of the associated secondary parameters that were determined to be invalid during the QC tests. No 5-minute value for a primary parameter was valid unless the following criteria were met:

1. The values of all the 5-minute averages for all of the associated secondary parameters had to be available.

2. The 5-minute values of all of the associated secondary parameters had to be valid.

Figure 5-6 lists journal entries for the time period corresponding to the preceding data listings. The journal entries were useful in arriving at decisions on the validity of flagged data.

A data review, illustrated in Figure 5-7 was produced for a quick overview of 5-minute average status for each data period. Each hour in this data listing is represented by 12 status bits,

CHAMP DATA, VALIDATED ON 01-AUG-79 WITH VERSION 7.00 OF VALDAT

VALIDATION REPORT FOR STATION - 0841 FOR DAY - 265 - 1977

HOURLY AVERAGES

TIME	NO_x	NO	NO_2	SNO_2	O_3	SO_2	CH_4	THC	NMHC	WS
121 0	0.0612 12	0.0069 12	0.0542 12	0.0568 12	0.1300 12	M 0	1.92 12	1.88 12	-0.02 12	7.7 12
131 0	0.0458 12	0.0044 12	0.0414 12	0.0437 12	0.1356 12	M 0	1.82 12	1.78 12	-0.02 12	8.5 12
141 0	0.0452M10	0.0045M10	0.0407M10	0.0431M10	0.1532B 4	M 0	1.84M10	1.80M10	-0.03M10	9.0M10
151 0	0.0402M 1	0.0039M 1	0.0363M 1	0.0386M 1	0.1574M 1	M 0	1.74M 1	1.70M 1	-0.02M 1	6.8M 1
161 0	0.0641M 6	0.0052M 6	0.0589M 6	0.0618M 6	0.1739M 6	M 0	1.94M 6	1.90M 6	-0.02M 6	11.5M 6

TIME	VMH	VWD	T_{OUT}	T_{IN}	DEWP	BP	HVFL	HSPI	HSPF
121 0	6.0 12	252.0 12	23.0 12	22.3 12	10.0 12	734.7 12	1.135 12	1.110 12	1.109 12
131 0	6.5 12	244.7 12	24.4 12	22.5 12	8.2 12	734.1 12	1.134 12	1.112 12	1.115 12
141 0	7.2M10	234.9M10	25.2M10	23.4M10	7.9M10	733.6M10	1.133M10	1.120M10	1.122M10
151 0	6.8M 1	309.4M 1	25.8M 1	23.4M 1	6.9M 1	733.5M 1	1.133M 1	M 0	M 0
161 0	11.2M 6	239.3M 6	25.0M 6	22.5M 6	10.3M 6	733.4M 6	1.134M 6	1.134M 6	1.139M 6

Figure 5-2. Example CHAMP data validation report (partial printout).

STATION - 0841 DAY 265-1977

VALIDATION CRITERIA USED

NO_2	VAC	LOW	LIMIT = 115.000	HIGH LIMIT = 135.000
SNO	FO_2	LOW	LIMIT = 28.000	HIGH LIMIT = 38.000
SNO	SFNO	LOW	LIMIT = 130.000	HIGH LIMIT = 300.000
SNO	VAC	LOW	LIMIT = 115.000	HIGH LIMIT = 135.000
→ O_3	FETH	LOW	LIMIT = 20.000	HIGH LIMIT = 30.000
O_3	SFO_3	LOW	LIMIT = 800.000	HIGH LIMIT =1600.000
SO_2	HSO_2	LOW	LIMIT = 165.000	HIGH LIMIT = 215.000
SO_2	SFSO	LOW	LIMIT = 130.000	HIGH LIMIT = 200.000

Figure 5-3. High/low critical values for CHAMP secondary parameters (partial printout).

STATION - 0841 DAY 265-1977

INVALIDITY CAUSES BY HOUR

O_3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14 FETH FTH/ SF	15	16	17	18	19	20	21	22	23
RSPI	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15 RSPI NEGV	16	17	18	19	20	21	22	23
RSPF	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15 RSPF NEGV	16	17	18	19	20	21	22	23

Figure 5-4. CHAMP validation system, invalidity causes by hour.

STATION - 0841 DAY 265-1977

INVALID SECONDARIES-FIVE MINUTE AVERAGES

14: 7	O_3 = 0.1449	FETH= 17.692	14:12	O_3 = 0.0626	FETH= -0.774	14:17	O_3 = 0.0182	FETH = -2.327
14:22	O_3 = 0.0102	FETH= -2.688	14:27	O_3 = 0.0147	FETH= 0.814	15: 5	RSPI= -0.924	RSPI= -0.924
15: 5	RSPF= -0.999	RSPF= -0.999						

Figure 5-5. Five-minute values of invalid secondary parameters.

STATION - 0841 DAY 265-1977

JOURNAL ENTRIES

STATION - 0841 DAY - 265 TIME - 13:45
WORD 9 HIT 1 GOES TO 1 STATE SOMETIMES: NO APPARENT REASON, R
NEPHELOMETER POWER IS NOT OFF. RICK F ON

STATION - 0841 DAY - 265 TIME - 13:48
OPC/PHA: WAS AT 50000 SECONDS & COUNTING, REFUSED TO TRANSFER R
DATA TO COMPUTER. USED A TO RESTART. RICK F E

STATION - 0841 DAY - 265 TIME - 15: 1
DAS: WAS PRINTING 7242, 5246, 0323 DURING POLL ATTEMPTS, R
NO RECORDS TAKEN, ZEROED CORE, LOADED PROGRAMS & CONSTANTS. R
RICK F D

STATION - 0841 DAY - 265 TIME - 16:31
DAS CRUSHED E 0002: DURING POLL ATTEMPT. TTY PRINTED 0316, R
0316, THEN TURNED OFF DIGITL DISPLAY, THEN WENT TO E 0002.
ZEROED CORE, LOADED PROGRAMS & CONSTANTS. RICK F E

STATION - 0841 DAY - 265 TIME - 16:46
TSP, RSP, ANDENSEN FILTERS CHANGED, SO_2 BUBBLER RUNNING: ANALYZER
B.O. EAA, OPC/PHA CALIBRATION ABORTED: DAS CRASHEDS, WARM, A
LITE HAZE TODAY. RICK F OI

STATION - 0841 DAY - 265 TIME - 22:22
DAVID TORRES AT GLENDORA TO CHECK OUT MODEM PROBLEMS

Figure 5-6. Example CHAMP journal entries for data validation.

Station - ∿841 Day 265-1977

****DATA REVIEW****

NO_x	0 0000000000000	1 0000000000000	2 0000000000000	3 0000000000000	4 0000000000000	5 0000000000000	6 0000000000000	7 0000000000
	8 0000000000000	9 0000000000000	10 0000000000000	11 0000000000000	12 0000000000000	13 0000000000000	14 0000000000--	15 -0--------
	16 ------000000	17 0000000000000	18 0000000000000	19 0000000000000	20 0000000000000	21 0000000000000	22 0000000000000	23 0000000000
NO	0 0000000000000	1 0000000000000	2 0000000000000	3 0000000000000	4 0000000000000	5 0000000000000	6 0000000000000	7 0000000000
	8 0000000000000	9 0000000000000	10 0000000000000	11 0000000000000	12 0000000000000	13 0000000000000	14 0000000000--	15 -0--------
	16 ------000000	17 0000000000000	18 0000000000000	19 0000000000000	20 0000000000000	21 0000000000000	22 0000000000000	23 0000000000
NO_2	0 0000000000000	1 0000000000000	2 0000000000000	3 0000000000000	4 0000000000000	5 0000000000000	6 0000000000000	7 0000000000
	8 0000000000000	9 0000000000000	10 0000000000000	11 0000000000000	12 0000000000000	13 0000000000000	14 0000000000--	15 -0--------
	16 ------000000	17 0000000000000	18 0000000000000	19 0000000000000	20 0000000000000	21 0000000000000	22 0000000000000	23 0000000000
SNO	0 0000000000000	1 0000000000000	2 0000000000000	3 0000000000000	4 0000000000000	5 0000000000000	6 0000000000000	7 0000000000
	8 0000000000000	9 0000000000000	10 0000000000000	11 0000000000000	12 0000000000000	13 0000000000000	14 0000000000--	15 -0--------
	16 ------000000	17 0000000000000	18 0000000000000	19 0000000000000	20 0000000000000	21 0000000000000	22 0000000000000	23 0000000000
O_3	0 0000000000000	1 0000000000000	2 0000000000000	3 0000000000000	4 0000000000000	5 0000000000000	6 0000000000000	7 0000000000
	8 0000000000000	9 0000000000000	10 0000000000000	11 0000000000000	12 0000000000000	13 0000000000000	14 01111!11000--	15 -0--------
	16 ------000000	17 0000000000000	18 0000000000000	19 0000000000000	20 0000000000000	21 0000000000000	22 0000000000000	23 0000000000
SO_2	0 ----------	1 ----------	2 ----------	3 ----------	4 ----------	5 ----------	6 ----------	7 ----------
	8 ----------	9 ----------	10 ----------	11 ----------	12 ----------	13 ----------	14 ----------	15 ----------
	16 ----------	17 ----------	18 ----------	19 ----------	20 ----------	21 ----------	22 ----------	23 ----------
CH_4	0 0000000000000	1 0000000000000	2 0000000000000	3 0000000000000	4 0000000000000	5 0000000000000	6 0000000000000	7 0000000000

Figure 5-7. Example CHMP validation data review.

each representing a 5-minute time interval. The 5-minute interval flags are represented symbolically as follows:

"O" - acceptable data
"I" - invalid data
"-" - missing data

Note that the flags for ozone data during hours 14 and 15 can be traced through the previous illustrations. For the cases in which 5-minute averages were missing, operator logs and strip charts were reviewed to determine if data could be captured; in cases where acceptable data were available, the data were entered into the data base.

Another feature of the CHAMP data validation routine was the application of a graphic technique. Figure 5-8 shows a plot of ozone as a function of time for day 265, the same period used in the preceding illustrations. The valid 5-minute averages are represented as O's; invalid 5-minute averages are entered as "I's". (The stacking of 5-minute values is merely a function of printer limitations). Better resolution was obtained in some cases by using a continuous line plotter. The illustration shows, however, that a useful graphic technique can be applied (relatively inexpensively) that flows well with other data reports produced for a specific time period.

The graphic display technique was useful in two ways--first, for quickly spotting extreme data values that warranted further investigation and secondly, for tracking hourly pollutant trends daily. In the case of ozone, the trend is clear in Figure 5-8.

Figure 5-9 is a complement to the ozone pattern, and it confirms that NO_x trends during the same day behaved in the way that would be expected from the photochemical relationships involved.

After the validation procedure was completed for CHAMP data, the data were approved for incorporation into the reporting data base. The automated and manual QC and data validation procedures in the CHAMP system have lent significant credibility to the final reported data.

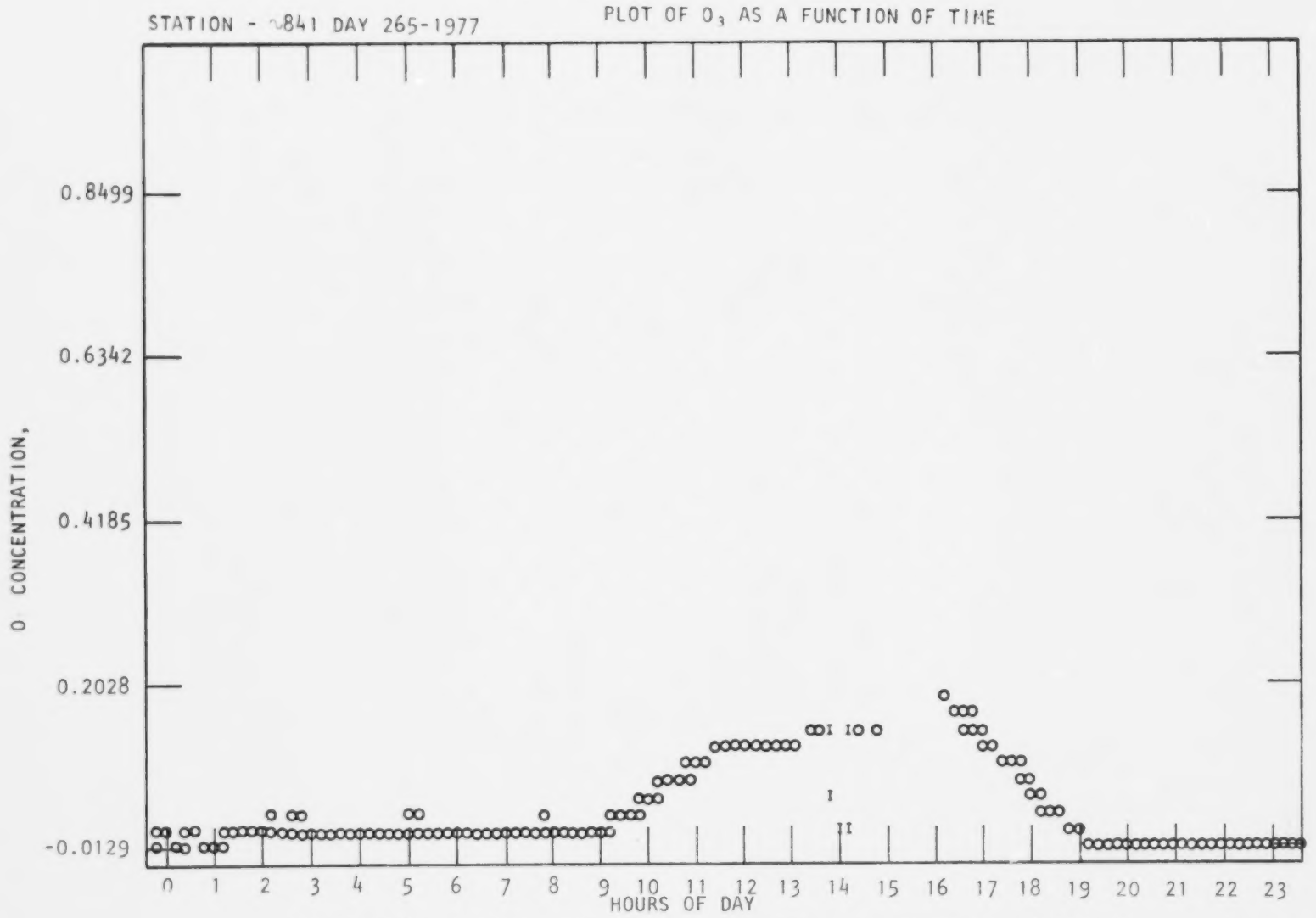


Figure 5.8. Example ozone daily data curve from CHAMP data validation.

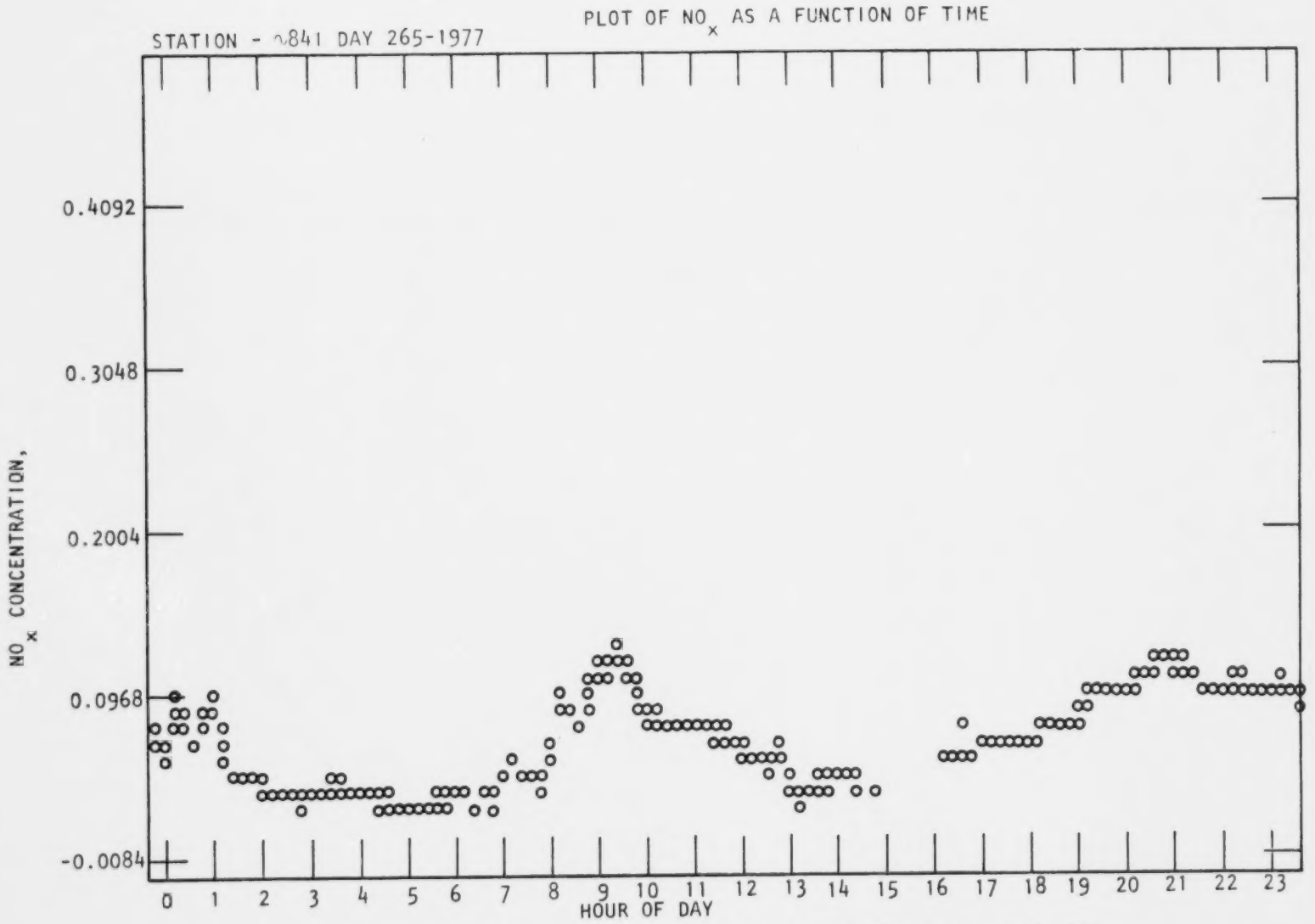


Figure 5.9. Example NO_x curve from CHAMP data validation.

5.5 CASE STUDY OF A REGIONAL AIR MONITORING SYSTEM (RAMS) DATA VALIDATION

The U.S. EPA operated the Regional Air Monitoring System (RAMS) in and around the St. Louis metropolitan area from 1975 through mid-1977. The system was part of the Regional Air Pollution Study (RAPS), undertaken to collect aerometric data for urban and rural scale dispersion model development and evaluation. The RAMS included 25 monitoring sites at which the parameters shown in Table 5-2 were monitored. Data were transmitted to a central computer facility from each monitoring site via leased lines. Figure 5-10 shows the relative locations of monitoring stations in the RAMS network. Figure 5-11 shows the flow of data within each station. In the RAMS system, minicomputers served QC and data handling functions similar to those in the CHAMP network. Unlike the CHAMP data base, which was developed with backup data tapes from each station, the RAMS data base was produced from data transmitted directly to the central computer. The station tapes in RAMS served as backup in instances when the telemetry system or central computer malfunctioned. Station data buffers were polled by the central computer every minute.

5.5.1 Quality Control

Several automated QC checks (based on the operating status of each instrument) were incorporated in RAMS to prescreen the data. The automated QC checks included:

- System status checks,
- Analog checks, and
- Zero/span checks.

The system status checks included approximately 35 electrical checks of critical parameters (e.g., flame-out, valve on-off status) to determine the capability of each instrument to produce valid data. When status checks indicated malfunctions, a flag was appended to the value of the associated environmental parameter to indicate that the 1-minute value was invalid.[a]

---

[a] System programmers developing programs to summarize data must be aware of the extent to which flags are used. In RAMS, the sample value was multiplied by 10^{-25} when a status condition was invalid. Problems arose when data analysis programs recognized such values as zero.

TABLE 5-2. PARAMETERS MONITORED IN THE REGIONAL AIR MONITORING SYSTEM

		Measurement interval (min)	Number of stations
Air quality:	Sulfur dioxide	3.75	13
	Total sulfur	1	12
	Hydrogen sulfide	3.75	13
	Ozone	1	25
	Nitric oxide	1	25
	Oxide of nitrogen	1	25
	Nitrogen dioxide	1	25
	Carbon monoxide	5	25
	Methane	5	25
	Total hydrocarbons	5	25
Meteorological:	Windspeed	1	25
	Wind direction	1	25
	Temperature	1	25
	Temperature gradient	1	7
	Pressure	1	7
	Dewpoint	1	25
	Aerosol scatter	1	25
Solar radiation:	Pyranometer	1	6
	Pyrheliometer	1	4
	Pyrgeometer	1	4

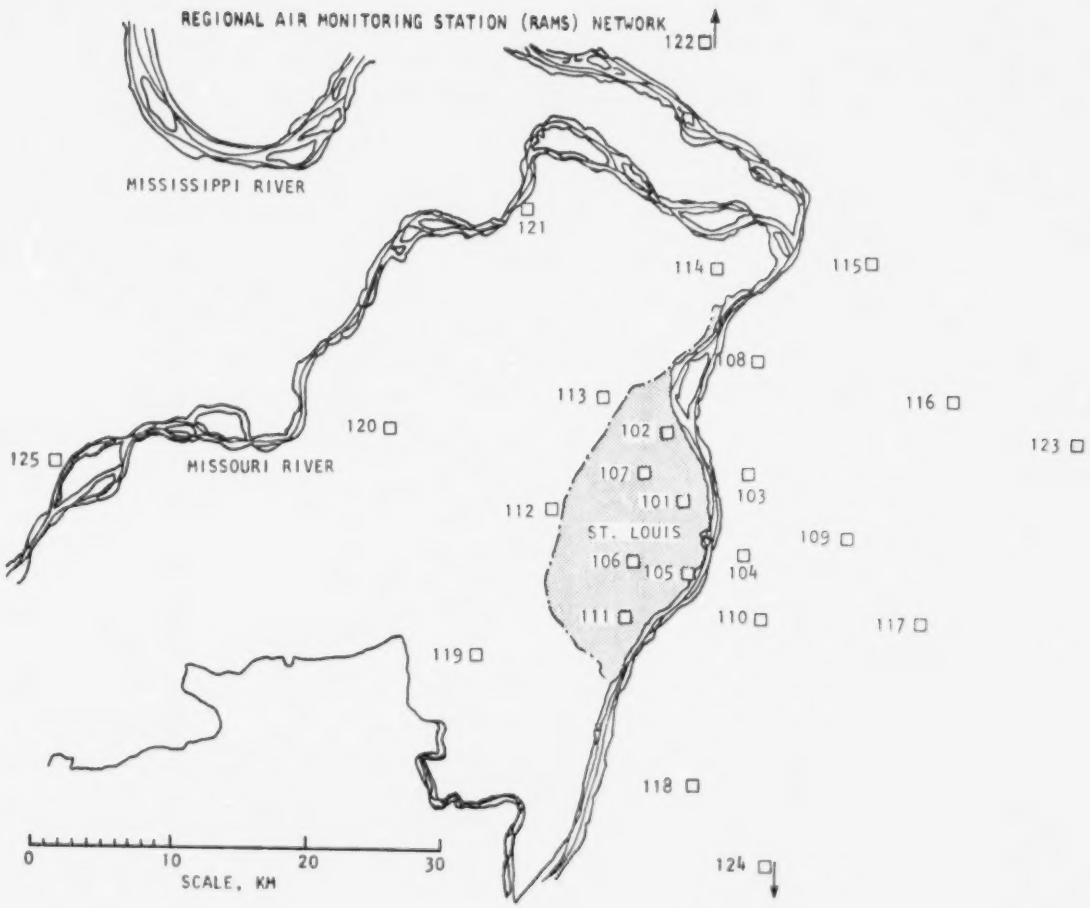


Figure 5-10. Location of RAMS stations.

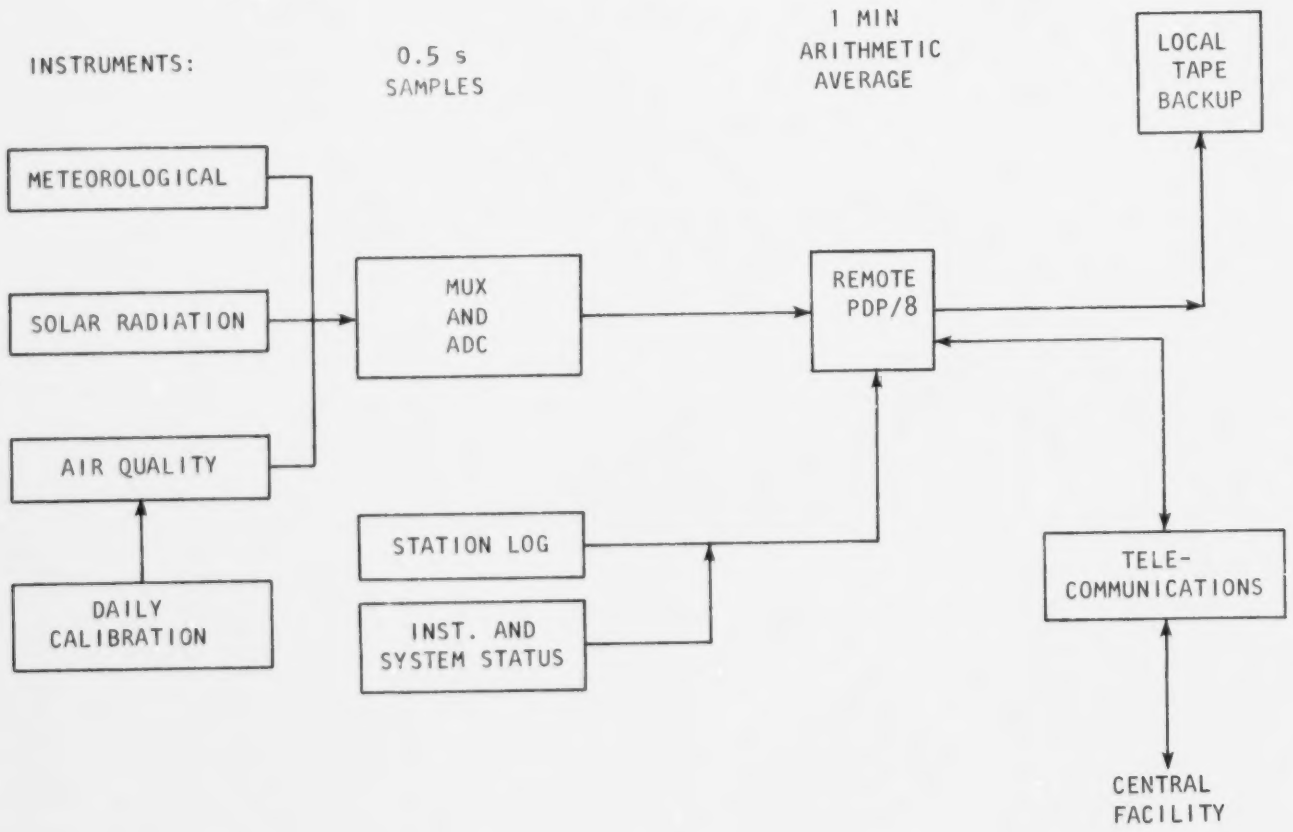


Figure 5-11. RAMS data flow: RAMS station.

Analog checks determined the status of several key parameter conditions such as permeation tube bath temperature and reference voltages. If acceptable limits for these parameters were exceeded, then the corresponding environmental data were invalidated. In this case, a value of 10^{34} was substituted for the 1-minute reading of the environmental parameter in the data file.

Zero/span checks were performed daily, with the zero/span commands coming from the central computer. If predetermined instrument drift limits were exceeded, the data from that instrument for the previous day were flagged as invalid, and a field operator was notified.

5.5.2 Data Validation

Figure 5-12 shows the data flow within the RAMS central facility. As the figure indicates, RAMS validation included two levels of data checks:

- Intrastation checks of minute data, and
- Interstation checks of hourly data.

The RAMS minute values that were not invalidated by the QC checks mentioned previously were converted to engineering units and then subjected to the intrastation checks. The automated intrastation checks included value limit checks, relational checks, and 10-minute time continuity checks. The gross limit checks were based on the operating ranges of each instrument. Examples of typical limits are in Table 5-3. Using the operating ranges avoided the problem of setting statistical limits for each parameter.

For some parameters, interparameter conditions were set as validation criteria; for example, total sulfur had to be less than the sum of SO_2 and hydrogen sulfide (when compared on a sulfur basis in ppm). Instrument noise bands precluded strict interpretation of some of the interparameter checks.

Ten-minute continuity checks were used to see if a variable changed over a period of time. It was recognized early in the RAMS program that a constant voltage output from a sensor indicated mechanical or electrical failures in the sensor instrumentation. A daily report was generated to show questionable time

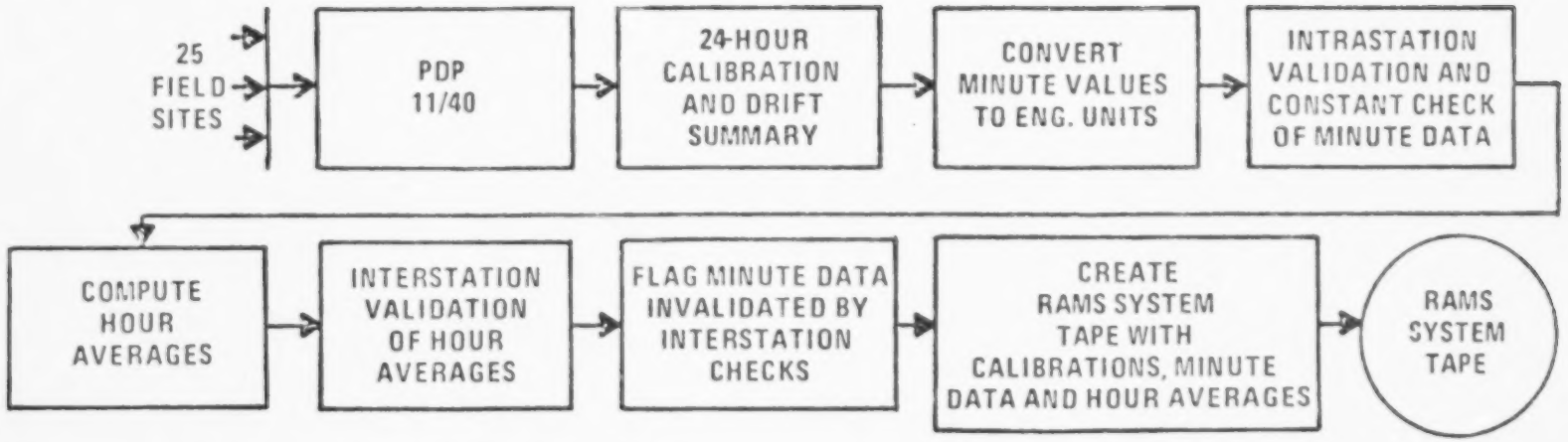


Figure 5-12. RAMS data flow, central facility.

TABLE 5-3. TYPICAL GROSS LIMIT VALUES USED IN THE REGIONAL AIR MONITORING SYSTEM DATA VALIDATION PROCEDURE

Parameter	Instrumental or natural limits		Interparameter condition
	Lower	Upper	
Ozone	0 ppm	1 ppm	$NO \times O_3 \leq 0.04$
Nitric oxide	0 ppm	5 ppm	$NO - NO_x \leq$ Noise (NO)
Oxides of nitrogen	0 ppm	5 ppm	$NO - NO_x \leq$ Noise (NO_x)
Carbon monoxide	0 ppm	100 ppm	
Methane	0 ppm	25 ppm	CH_4 - THC $\leq$ Noise (CH_4)
Total hydrocarbons	0 ppm	25 ppm	CH_4 - THC $\leq$ Noise (THC)
Sulfur dioxide	0 ppm	1 ppm	SO_2 - TS $\leq$ Noise (SO_2)
Total sulfur	0 ppm	1 ppm	SO_2 - TS $\leq$ Noise (TS)
Hydrogen sulfide	0 ppm	1 ppm	H_2S - TS $\leq$ Noise (H_2S)
Aerosol scatter	$0.000001 m^{-1}$	$0.0040 m^{-1}$	
Windspeed	0 m/s	22.2 m/s	
Wind direction	0°	360° (540° for some wind systems)	
Temperature	-20°C	45°C	
Dewpoint	-30°C	45°C	DP - 0.5 $\leq$ T
Temperature gradient	-5°C	5°C	
Barometric pressure	950 mb	1050 mb	
Pyranometers	-0.50	2.50 Langleys/min	
Pyrgeometers	0.30	0.75 Langleys/min	
Pyrheliometers	-0.50	2.50 Langleys/min	

periods so that field personnel could be dispatched to a site if instrument malfunctions were suspected. These checks could not be applied to some parameters; for example, barometric pressure, since it can remain constant (to the number of digits recorded) much longer than ten minutes.

The interstation checks were applied to hourly averages compiled from the minute data that passed both the on-site QC tests and the after-the-fact intrastation checks. Interstation continuity checks were used to look for consistency in parameter behavior throughout the network during specific time periods. The interstation checks in RAMS were applied to meteorological parameters only. Initial application to pollutant parameters showed that too many false data anomalies were flagged. The interstation checks consisted of plotting curves of hourly averages of each parameter. Several curves representing data from multiple sites were plotted on the same page for ease of comparison among stations. The mean and difference of values that fell outside the 90th percentile of the combined network distribution were reported. Data curves were reviewed over periods of several days to see trend relationships to the rest of the network. When problems were detected through the review of the hourly average curves, then plots of the minute data for the corresponding time periods were produced. Figure 5-13 is a continuous plot of hourly average temperature data for several stations in the RAMS network. The hourly data were relatively uniform among the nine stations represented; however, the small negative depression on the circled portion of the curve for Station 103 on the thirty-second Julian day. Figure 5-14, a plot of minute temperature data for the same station and time period, reveals that a problem occurred during most of hour 14 of the day in question and that the data reported for that hour are invalid.

This example shows that, although hourly average data may not have indicated instrument problems over the long term, problems such as temporary voltage surges may indeed have occurred. In the example, statistical analysis of hourly average values would not have detected the problem. In the short term this kind

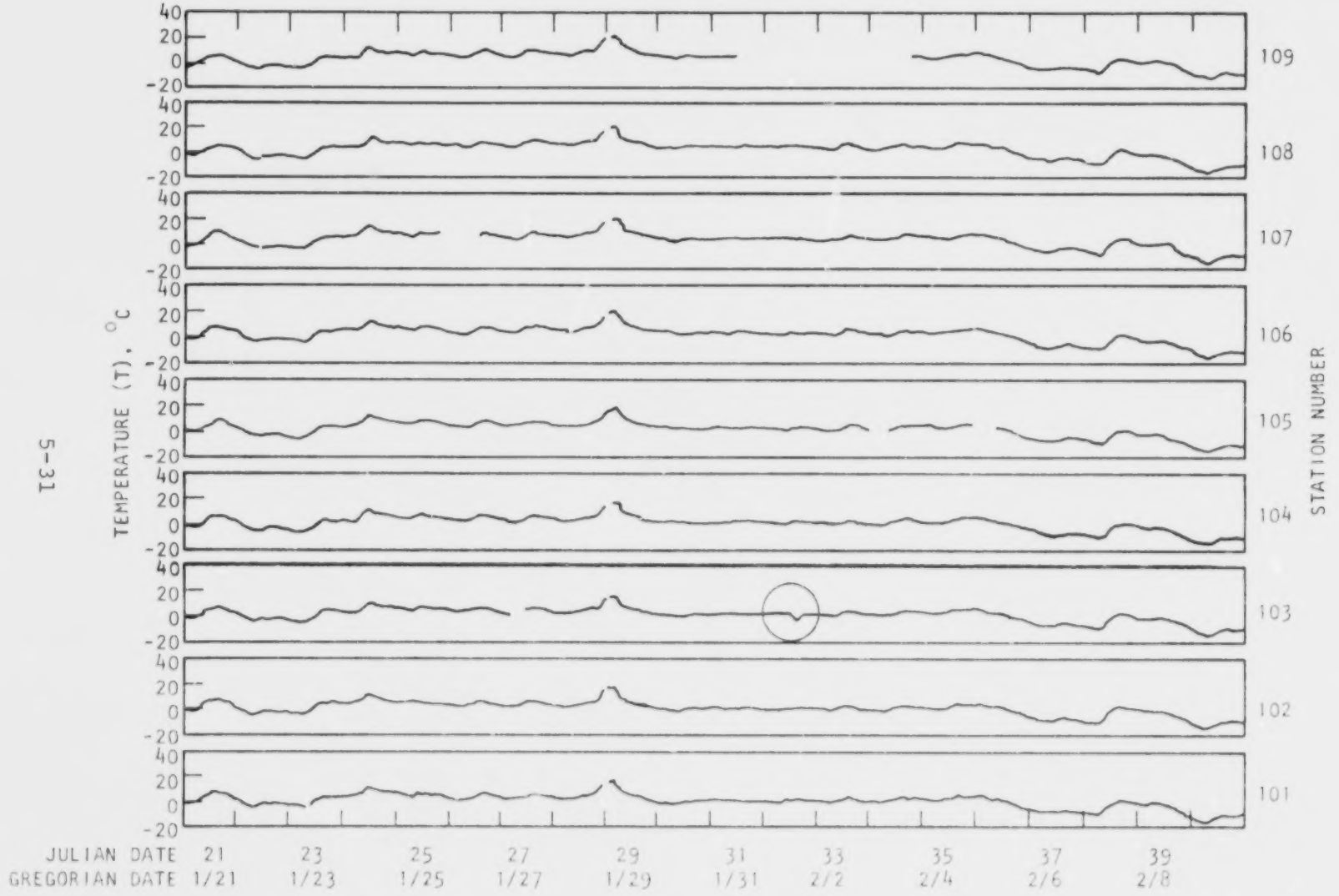


Figure 5-13. RAMS hourly average temperature data, January 21, to February 9, 1975.

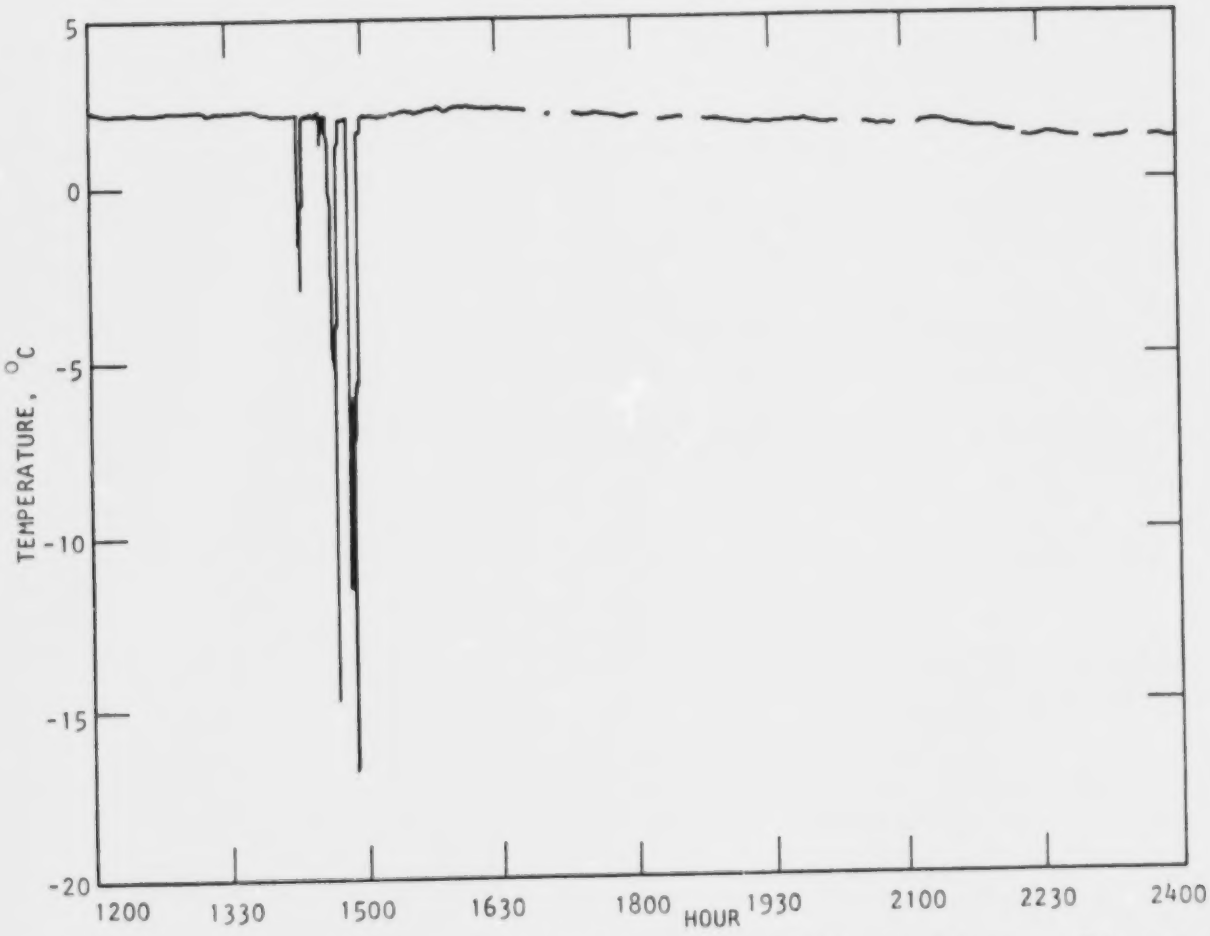


Figure 5-14. Minute temperature data from Station 103, from 1200 to 2359 hours, February 1, 1975.

of information may not be important for hourly or daily data fluctuations, but it can be important as a quality assurance feedback mechanism. Even though the QC checks were extensive, they were not sufficient to detect this problem. The validation procedure was a supplement to the QC procedures that may be useful in improving those procedures as well as in identifying additional instrument function problems.

The use of data plotting techniques for hourly and minute data coupled with followup to field operation loop was helpful in answering questions regarding data values flagged during the intrastation validation tests. A special data listing was produced to show the average standard deviation and the number of excursions above statistical limits. The largest deviations were flagged for followup. This technique was extremely time consuming, since the followup was manual.

Meteorological data were used in RAMS to provide insight to questionable data periods. Weather summaries like the one in Figure 5-15 provided convenient references for determining conditions that might have significantly altered pollutant patterns during specific periods of time.

In retrospect, two specific changes in approach might have improved the validation effort in RAMS:

1. Graphic techniques such as those described herein could have been more useful if they had been applied closer to real time. The major problem is that (as mentioned previously) the graphical techniques are manpower intensive, so condensing the time frame (especially for a network approaching RAMS size) could markedly increase the validation cost.

2. Successive difference checks could have been applied to minute (or hourly average) data to provide feedback to the QC system. These checks could solve the problem of an instrument malfunction going undetected for a long period of time. Figure 5-16 illustrates the kinds of instrument-response problems that might be detected using successive difference checks of minute (or hourly average) data. Optimum limits need to be determined for the successive difference checks by a statistical analysis of

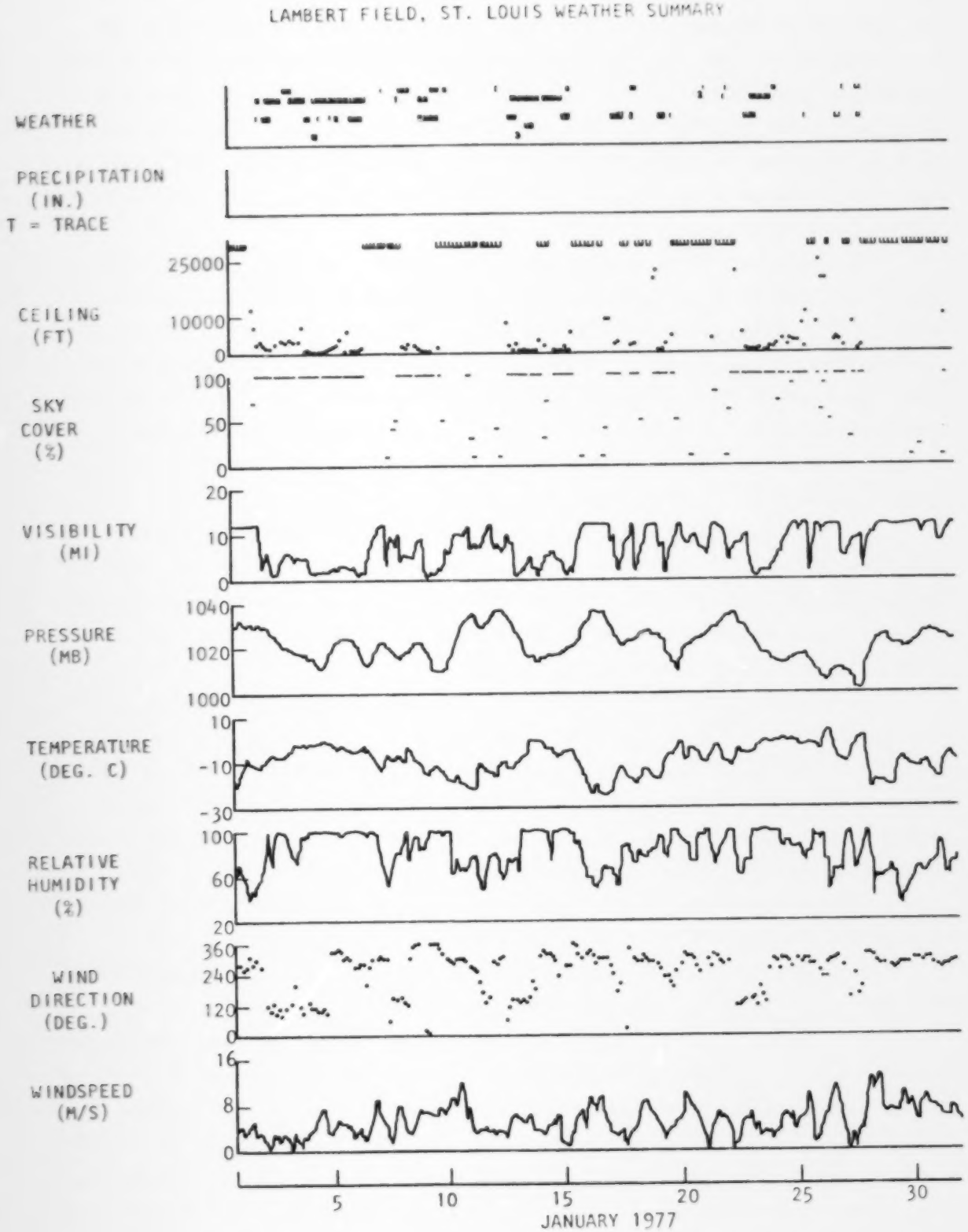


Figure 5-15. Reference weather data used in RAMS data validation.

historical data. For some applications it will be necessary to determine successive differences on a percentage basis (i.e., above a specified level).[3]

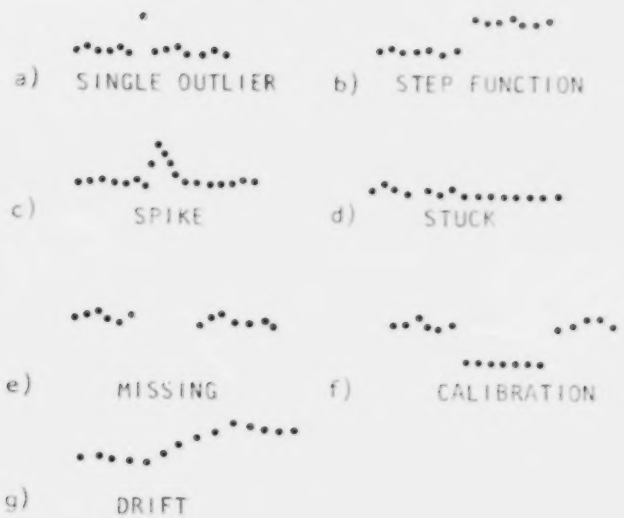


Figure 5-16. Examples of instrument responses that can be detected through minute successive differences.

5.6 REFERENCES

1. U.S. Environmental Protection Agency. Quality Assurance Handbook. Vol. I. Principles. EPA-600/9-76-005, 1976.

2. Rhodes, R. C., and S. Hochheiser. Data Validation Conference Proceedings. Presented by Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, EPA 600/9-79-042, September 1979.

3. Hartwell, T., Use of Successive Time Difference and Dixon Ratio Test for Data Validation, Data Validation Conference Proceedings, EPA-600/9-79-042, September 1979.

6.0 BIBLIOGRAPHY

Barnett, V., and T. Lewis. Outliers in Statistical Data. John Wiley and Sons, New York, 1978.

Curran, T. C., W. F. Hunt, Jr., and R. B. Faoro. Quality Control for Hourly Air Pollution Data. Presented at the 31st Annual Technical Conference of the American Society for Quality Control, Philadelphia, May 16-18, 1977.

Data Validation Program for SAROAD, Northrup Services, EST-TN-78-09, December 1978, (also see Program Documentation Manual, EMSL).

Faoro, R. B., T. C. Curran, and W. F. Hunt, Jr., "Automated Screening of Hourly Air Quality Data," Transactions of the American Society for Quality Control, Chicago, Ill., May 1978.

Grant, E. L., and R. S. Leavenworth. Statistical Quality Control. McGraw-Hill Book Company, New York.

Grubbs, F. E., and G. Beck. Extension of Sample Sizes and Percentage Points for Significance Test of Outlying Observations. Technometrics, Vol. 14, No. 4, November 1972.

Hald, A. Statistical Theory with Engineering Applications. New York, 1952.

Hartwell, T., Use of Successive Time Difference and Dixon Ratio Test for Data Validation, Data Validation Conference Proceedings, EPA-600/9-79-042, September 1979.

Hawkins, D. M. The Detection of Errors in Multivariate Data Using Principal Components. Journal of the American Statistical Association, Vol. 69, No. 346. 1974.

Hunt, Jr., W. F., J. B. Clark, and S. K. Goranson, "The Shewhart Control Chart Test: A Recommended Procedure for Screening 24-Hour Air Pollution Measurements," J. Air Poll. Control Assoc. 28:508 (1979).

Hunt, Jr., W. F., T. C. Curran, N. H. Frank, and R. B. Faoro, "Use of Statistical Quality Control Procedures in Achieving and Maintaining Clean Air," Transactions of the Joint European Organization for Quality Control/International Academy for Quality Conference, Vernice Lido, Italy, September 1975.

Hunt, Jr., W. F., R. B. Faoro, T. C. Curran, and W. M. Cox, "The Application of Quality Control Procedures to the Ambient Air Pollution Problem in the USA," Transactions of the European Organization for Quality Control, Copenhagen, Denmark, June 1976.

Hunt, Jr., W. F., R. B. Faoro, and S. K. Goranson, "A Comparison of the Dixon Ratio Test and Shewhart Control Test Applied to the National Aerometric Data Bank," Transactions of the American Society for Quality Control, Toronto, Canada, June 1976.

Johnson, T. A Comparison of the Two-Parameter Weibull and Log-normal Distributions Fitted to Ambient Ozone Data. PEDCo Environmental, Inc., Durham, North Carolina, and The Air Pollution Control Association. Quality Assurance in Air Pollution Measurement. Presented at the Air Pollution Control Association, New Orleans, March 11-14, 1979.

Larsen, R. I. A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards, Publication No. AP-89, U.S. Environmental Protection Agency, 1971.

Marriott, F. H. C. The Interpretation of Multiple Observations. Academic Press, New York, 1974.

Naus, J. I. Data Quality Control and Editing. Marcel Dekker, Inc., New York, 1975.

Remington, R. D., and M. A. Schork. Statistics with Applications to the Biological and Health Sciences. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970.

Rhodes, R. C., and S. Hochheiser. Data Validation Conference Proceedings. Presented by Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, EPA 600/9-79-042, September 1979.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, 1956.

Smith, F. Guideline for the Development and Implementation of a Quality Cost System for Air Pollution Measurement Programs. Research Triangle Institute. RTI/1507/01F, November 1979.

U.S. Department of Commerce. Computer Science and Technology: Performance Assurance and Data Integrity Practices. National Bureau of Standards, Washington, D.C., January 1978.

U.S. Environmental Protection Agency. Guidelines for Air Quality Maintenance Planning and Analysis. Vol. 11. Air Quality Monitoring and Data Analysis. EPA-450/4-74-012, 1974.

U.S. Environmental Protection Agency. Quality Assurance and Data Validation for the Regional Air Monitoring System of the St. Louis Regional Air Pollution Study. EPA-600/4-76-016, 1976.

U.S. Environmental Protection Agency. Quality Assurrance Handbook. Vol. I. Principles. EPA-600/9-76-005, 1976.

U.S. Environmental Protection Agency. Quality Assurance Handbook: Vol. I, Principles; Vol. II, Ambient Air Specific Methods; and Vol. III, Stationary Source Specific Methods. EPA-600/9-76-005, 1976.

U.S. Environmental Protection Agency. Screening Procedures for Ambient Air Quality Data. EPA-450/2-78-037, July 1978.

Wolfe, J. H. NORMIX: Computation Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions, Research Memorandum SRM 68-2. U.S. Naval Personnel Research Activity, San Diego, 1967.

1978 Annual Book of ASTM Standards, Part 41. Standard Recommended Practice for Dealing with Outlying Observations, ASTM Designation: E 178-75. pp. 212-240.

APPENDIX A

STATISTICAL TABLES

TABLE A-1. DIXON CRITERIA FOR TESTING OF EXTREME OBSERVATION (SINGLE SAMPLE)*

n	Criterion	Significance level 10%	5%	1%
3	$r_{10} = \frac{x_2 - x_1}{x_n - x_1}$ if* smallest value is suspected; $= \frac{x_n - x_{n-1}}{x_n - x_1}$ if largest value is suspected.	.886	.941	.988
4		.679	.765	.889
5		.557	.642	.780
6		.482	.560	.698
7		.434	.507	.637
8	$r_{11} = \frac{x_2 - x_1}{x_{n-1} - x_1}$ if smallest value is suspected; $= \frac{x_n - x_{n-1}}{x_n - x_2}$ if largest value is suspected.	.479	.554	.683
9		.441	.512	.635
10		.409	.447	.597
11	$r_{31} = \frac{x_3 - x_1}{x_{n-1} - x_1}$ if smallest value is suspected. $= \frac{x_n - x_{n-2}}{x_n - x_2}$ if largest value is suspected.	.517	.576	.679
12		.490	.546	.642
13		.467	.521	.615
14	$r_{22} = \frac{x_3 - x_1}{x_{n-2} - x_1}$ if smallest value is suspected. $= \frac{x_n - x_{n-2}}{x_n - x_3}$ if largest value is suspected;	.492	.546	.641
15		.472	.525	.616
16		.454	.507	.595
17		.438	.490	.577
18		.424	.475	.561
19		.412	.462	.547
20		.401	.450	.535
21		.391	.440	.524
22		.382	.430	.514
23		.374	.421	.505
24		.367	.413	.497
25		.360	.406	.489

*Reproduced with permission from W. J. Dixon, "Processing Data for Outliers," Biometrics, March 1953, Vol. 9, No. 1, Appendix, Page 89.

$x_1 \leq x_2 \leq \cdots \leq x_{n-2} \leq x_{n-1} \leq x_n$

Criterion r_{10} applies for $3 \leq n \leq 7$

Criterion r_{11} applies for $8 \leq n \leq 10$

Criterion r_{21} applies for $11 \leq n \leq 13$

Criterion r_{22} applies for $14 \leq n \leq 25$

TABLE A-2. CRITICAL VALUES FOR 5% AND 1% TESTS OF DISCORDANCY FOR TWO OUTLIERS IN A NORMAL SAMPLE*

n	5%	1%
4	0.967	0.992
5	0.845	0.929
6	0.736	0.836
7	0.661	0.778
8	0.607	0.710
9	0.565	0.667
10	0.531	0.632
12	0.481	0.579
14	0.445	0.538
16	0.418	0.508
18	0.397	0.484
20	0.372	0.464
25	0.343	0.428
30	0.322	0.402

*Barnett, V., and T. Lewis. Outliers in Statistical Data, Table XIIIe, p. 311. John Wiley and Sons, New York, 1978.

TABLE A-3. CRITICAL T VALUES FOR ONE-SIDED GRUBBS TEST WHEN STANDARD DEVIATION IS CALCULATED FROM SAMPLE*

Number of Observations n	Upper .1% Significance Level	Upper .5% Significance Level	Upper 1% Significance Level	Upper 2.5% Significance Level	Upper 5% Significance Level	Upper 10% Significance Level
3	1.155	1.155	1.155	1.155	1.153	1.148
4	1.499	1.496	1.492	1.481	1.463	1.425
5	1.780	1.764	1.749	1.715	1.672	1.602
6	2.011	1.973	1.944	1.887	1.822	1.729
7	2.201	2.139	2.097	2.020	1.938	1.828
8	2.358	2.274	2.221	2.126	2.032	1.909
9	2.492	2.387	2.323	2.215	2.110	1.977
10	2.606	2.482	2.410	2.290	2.176	2.036
11	2.705	2.564	2.485	2.355	2.234	2.088
12	2.791	2.636	2.550	2.412	2.285	2.134
13	2.867	2.699	2.607	2.462	2.331	2.175
14	2.935	2.755	2.659	2.507	2.371	2.213
15	2.997	2.806	2.705	2.549	2.409	2.247
16	3.052	2.852	2.747	2.585	2.443	2.279
17	3.103	2.894	2.785	2.620	2.475	2.309
18	3.149	2.932	2.821	2.651	2.504	2.335
19	3.191	2.968	2.854	2.681	2.532	2.361
20	3.230	3.001	2.884	2.709	2.557	2.385
21	3.266	3.031	2.912	2.733	2.580	2.408
22	3.300	3.060	2.939	2.758	2.603	2.429
23	3.332	3.087	2.963	2.781	2.624	2.448
24	3.362	3.112	2.987	2.802	2.644	2.467
25	3.389	3.135	3.009	2.822	2.663	2.486
26	3.415	3.157	3.029	2.841	2.681	2.502
27	3.440	3.178	3.049	2.859	2.698	2.519
28	3.464	3.199	3.068	2.876	2.714	2.534
29	3.486	3.218	3.085	2.893	2.730	2.549
30	3.507	3.236	3.103	2.908	2.745	2.563
31	3.528	3.253	3.119	2.924	2.759	2.577
32	3.548	3.270	3.135	2.938	2.773	2.591
33	3.565	3.286	3.150	2.952	2.786	2.604
34	3.582	3.301	3.164	2.965	2.799	2.616
35	3.599	3.316	3.178	2.979	2.811	2.628
36	3.616	3.330	3.191	2.991	2.823	2.639
37	3.631	3.343	3.204	3.003	2.835	2.650
38	3.646	3.356	3.216	3.014	2.846	2.661
39	3.660	3.369	3.228	3.025	2.857	2.671
40	3.673	3.381	3.240	3.036	2.866	2.682
41	3.687	3.393	3.251	3.046	2.8[illegible]	2.692
42	3.700	3.404	3.261	3.057	2.887	2.700
43	3.712	3.415	3.271	3.067	2.896	2.710
44	3.724	3.425	3.282	3.075	2.905	2.719
45	3.736	3.435	3.292	3.085	2.914	2.727
46	3.747	3.445	3.302	3.094	2.923	2.736
47	3.757	3.455	3.310	3.103	2.931	2.744
48	3.768	3.464	3.319	3.111	2.940	2.753
49	3.779	3.474	3.329	3.120	2.948	2.760
50	3.789	3.483	3.336	3.128	2.956	2.768

*Grubbs, F. E., and G. Beck. Extension of Sample Sizes and Percentage Points for Significance Test of Outlying Observations. Technometrics, Vol. 14, No. 4, November 1972.

(continued)

TABLE A-3 (continued)

Number of Observations n	Upper .1% Significance Level	Upper .5% Significance Level	Upper 1% Significance Level	Upper 2.5% Significance Level	Upper 5% Significance Level	Upper 10% Significance Level
51	3.798	3.491	3.345	3.136	2.964	2.775
52	3.808	3.500	3.353	3.143	2.971	2.783
53	3.816	3.507	3.361	3.151	2.978	2.790
54	3.825	3.516	3.368	3.158	2.986	2.798
55	3.834	3.524	3.376	3.166	2.992	2.804
56	3.842	3.531	3.383	3.172	3.000	2.811
57	3.851	3.539	3.391	3.180	3.006	2.818
58	3.858	3.546	3.397	3.186	3.013	2.824
59	3.867	3.553	3.405	3.193	3.019	2.831
60	3.874	3.560	3.411	3.199	3.025	2.837
61	3.882	3.566	3.418	3.205	3.032	2.842
62	3.889	3.573	3.424	3.212	3.037	2.849
63	3.896	3.579	3.430	3.218	3.044	2.854
64	3.903	3.586	3.437	3.224	3.049	2.860
65	3.910	3.592	3.442	3.230	3.055	2.866
66	3.917	3.598	3.449	3.235	3.061	2.871
67	3.923	3.605	3.454	3.241	3.066	2.877
68	3.930	3.610	3.460	3.246	3.071	2.883
69	3.936	3.617	3.466	3.252	3.076	2.888
70	3.942	3.622	3.471	3.257	3.082	2.893
71	3.948	3.627	3.476	3.262	3.087	2.897
72	3.954	3.633	3.482	3.267	3.092	2.903
73	3.960	3.638	3.487	3.272	3.098	2.908
74	3.965	3.643	3.492	3.278	3.102	2.912
75	3.971	3.648	3.496	3.282	3.107	2.917
76	3.977	3.654	3.502	3.287	3.111	2.922
77	3.982	3.658	3.507	3.291	3.117	2.927
78	3.987	3.663	3.511	3.297	3.121	2.931
79	3.992	3.669	3.516	3.301	3.125	2.935
80	3.998	3.673	3.521	3.305	3.130	2.940
81	4.002	3.677	3.525	3.309	3.134	2.945
82	4.007	3.682	3.529	3.315	3.139	2.949
83	4.012	3.687	3.534	3.319	3.143	2.953
84	4.017	3.691	3.539	3.323	3.147	2.957
85	4.021	3.695	3.543	3.327	3.151	2.961
86	4.026	3.699	3.547	3.331	3.155	2.966
87	4.031	3.704	3.551	3.335	3.160	2.970
88	4.035	3.708	3.555	3.339	3.163	2.973
89	4.039	3.712	3.559	3.343	3.167	2.977
90	4.044	3.716	3.563	3.347	3.171	2.981
91	4.049	3.720	3.567	3.350	3.174	2.984
92	4.053	3.725	3.570	3.355	3.179	2.989
93	4.057	3.728	3.575	3.358	3.182	2.993
94	4.060	3.732	3.579	3.362	3.186	2.996
95	4.064	3.736	3.582	3.365	3.189	3.000
96	4.069	3.739	3.586	3.369	3.193	3.003
97	4.073	3.744	3.589	3.372	3.196	3.006
98	4.076	3.747	3.593	3.377	3.201	3.011
99	4.080	3.750	3.597	3.380	3.204	3.014
100	4.084	3.754	3.600	3.383	3.207	3.017

(continued)

TABLE A-3 (continued)

Number of Observations n	Upper .1% Significance Level	Upper .5% Significance Level	Upper 1% Significance Level	Upper 2.5% Significance Level	Upper 5% Significance Level	Upper 10% Significance Level
101	4.088	3.757	3.603	3.386	3.210	3.021
102	4.092	3.760	3.607	3.390	3.214	3.024
103	4.095	3.765	3.610	3.393	3.217	3.027
104	4.098	3.768	3.614	3.397	3.220	3.030
105	4.102	3.771	3.617	3.400	3.224	3.033
106	4.105	3.774	3.620	3.403	3.227	3.037
107	4.109	3.777	3.623	3.406	3.230	3.040
108	4.112	3.780	3.626	3.409	3.233	3.043
109	4.116	3.784	3.629	3.412	3.236	3.046
110	4.119	3.787	3.632	3.415	3.239	3.049
111	4.122	3.790	3.636	3.418	3.242	3.052
112	4.125	3.793	3.639	3.422	3.245	3.055
113	4.129	3.796	3.642	3.424	3.248	3.058
114	4.132	3.799	3.645	3.427	3.251	3.061
115	4.135	3.802	3.647	3.430	3.254	3.064
116	4.138	3.805	3.650	3.433	3.257	3.067
117	4.141	3.808	3.653	3.435	3.259	3.070
118	4.144	3.811	3.656	3.438	3.262	3.073
119	4.146	3.814	3.659	3.441	3.265	3.075
120	4.150	3.817	3.662	3.444	3.267	3.078
121	4.153	3.819	3.665	3.447	3.270	3.081
122	4.156	3.822	3.667	3.450	3.274	3.083
123	4.159	3.824	3.670	3.452	3.276	3.086
124	4.161	3.827	3.672	3.455	3.279	3.089
125	4.164	3.831	3.675	3.457	3.281	3.092
126	4.166	3.833	3.677	3.460	3.284	3.095
127	4.169	3.836	3.680	3.462	3.286	3.097
128	4.173	3.838	3.683	3.465	3.289	3.100
129	4.175	3.840	3.686	3.467	3.291	3.102
130	4.178	3.843	3.688	3.470	3.294	3.104
131	4.180	3.845	3.690	3.473	3.296	3.107
132	4.183	3.848	3.693	3.475	3.298	3.109
133	4.185	3.850	3.695	3.478	3.302	3.112
134	4.188	3.853	3.697	3.480	3.304	3.114
135	4.190	3.856	3.700	3.482	3.306	3.116
136	4.193	3.858	3.702	3.484	3.309	3.119
137	4.196	3.860	3.704	3.487	3.311	3.122
138	4.198	3.863	3.707	3.489	3.313	3.124
139	4.200	3.865	3.710	3.491	3.315	3.126
140	4.203	3.867	3.712	3.493	3.318	3.129
141	4.205	3.869	3.714	3.497	3.320	3.131
142	4.207	3.871	3.716	3.499	3.322	3.133
143	4.209	3.874	3.719	3.501	3.324	3.135
144	4.212	3.876	3.721	3.503	3.326	3.138
145	4.214	3.879	3.723	3.505	3.328	3.140
146	4.216	3.881	3.725	3.507	3.331	3.142
147	4.219	3.883	3.727	3.509	3.334	3.144

TABLE A-4. WILCOXON SIGNED-RANK TEST*

n = number of pairs

Number of pairs, n	Critical values			
	$\alpha \leq 0.10$	$\alpha \leq 0.05$	$\alpha \leq 0.02$	$\alpha \leq 0.01$
1				
2				
3				
4				
5	0, 15			
6	2, 19	0, 21		
7	3, 25	2, 26	0, 28	
8	5, 31	3, 33	1, 35	0, 36
9	8, 37	5, 40	3, 42	1, 44
10	10, 45	8, 47	5, 50	3, 52
11	13, 53	10, 56	7, 59	5, 61
12	17, 61	13, 65	9, 69	7, 71
13	21, 70	17, 74	12, 79	9, 82
14	25, 80	21, 84	15, 90	12, 93
15	30, 90	25, 95	19, 101	15, 105
16	35, 101	29, 107	23, 113	19, 117
17	41, 112	34, 119	28, 125	23, 130
18	47, 124	40, 131	32, 139	27, 144
19	53, 137	46, 144	37, 153	33, 158
20	60, 150	52, 158	43, 167	37, 173
21	67, 164	58, 173	49, 182	42, 189
22	75, 178	66, 187	55, 198	48, 205
23	83, 193	73, 203	62, 214	54, 222
24	91, 209	81, 210	69, 231	61, 239
25	100, 225	89, 236	76, 249	68, 257

*The data of this table is reproduced from Documenta GEIGY, Scientific Tables 7th edition. With kind permission of CIBA-BEIGY Ltd., Basle (Switzerland).

TABLE A-5. RANK SUM TEST $\alpha = P[H_0 \text{ is true}]$*

In performing the test, begin with the $\alpha = 0.10$ table and if T_1 does not fall between T_1 and T_2, repeat the test using $\alpha = 0.05$, $\alpha = 0.02$, and $\alpha = 0.01$, until the inequality is satisfied.

$\alpha = 0.01$

N_2	4		5		6		7		8		9		10		11		12		13		14		15		16		17		18		19		20		21		22		23		24		25	
N_1	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
4	—	—	—	—	21-	45	28-	50	37-	67	48-	80	37-	93	68-	108	81-	133	96-	140	109-	157	125-	175	143-	195	159-	215	177-	237	197-	259	218-	282	239-	307	262-	332	285-	359	310-	386	335-	413
5	—	—	15-	40	22-	50	29-	63	38-	76	48-	87	59-	101	71-	116	84-	132	98-	149	113-	168	128-	187	145-	207	163-	228	182-	250	202-	273	223-	297	245-	321	267-	349	291-	376	316-	404	342-	433
6	10-	34	16-	44	23-	55	31-	67	40-	80	50-	94	61-	109	73-	123	87-	141	101-	159	116-	178	132-	198	149-	219	168-	240	187-	261	207-	287	228-	312	250-	338	274-	364	298-	392	325-	421	349-	451
7	10-	38	16-	49	24-	60	32-	73	42-	86	52-	101	64-	116	76-	133	90-	150	104-	169	120-	188	136-	209	154-	230	172-	253	192-	276	212-	301	234-	326	256-	[illegible]	280-	380	305-	408	330-	438	357-	468
8	11-	41	17-	53	25-	65	34-	78	45-	93	54-	106	66-	124	79-	141	93-	159	108-	178	123-	199	140-	220	158-	242	177-	265	197-	289	218-	314	240-	340	26	?	287-	393	312-	425	337-	455	364-	486
9	11-	45	18-	57	26-	70	33-	84	43-	99	56-	115	68-	133	82-	149	96-	168	111-	188	127-	209	144-	231	163-	253	182-	277	202-	302	223-	328	246-	354	269-	?	293-	411	319-	440	345-	471	372-	503
10	12-	48	19-	61	27-	75	37-	89	47-	103	58-	122	71-	139	86-	158	99-	177	115-	197	133-	219	149-	241	167-	265	187-	289	208-	314	229-	341	252-	368	275-	397	300-	426	326-	456	352-	488	380-	520
11	12-	52	20-	65	28-	80	38-	93	49-	111	61-	128	73-	147	87-	166	102-	189	118-	207	133-	229	153-	252	172-	276	192-	301	213-	327	235-	354	258-	382	282-	411	307-	442	333-	472	360-	504	388-	537
12	13-	53	21-	69	30-	84	40-	100	51-	117	63-	135	76-	154	90-	174	105-	195	122-	216	139-	239	157-	261	177-	287	197-	313	218-	340	241-	367	264-	396	289-	425	314-	456	340-	488	368-	520	396-	554
13	13-	39	22-	73	31-	89	41-	106	53-	123	65-	142	79-	161	93-	182	109-	203	125-	226	143-	249	162-	273	181-	299	202-	325	224-	352	247-	380	271-	409	295-	440	321-	471	348-	503	376-	536	404-	571
14	14-	62	23-	78	32-	94	43-	111	54-	130	67-	149	81-	169	96-	190	112-	212	129-	233	147-	259	166-	284	186-	310	208-	336	230-	364	253-	393	277-	423	302-	454	328-	486	355-	519	383-	553	413-	587
15	15-	65	23-	82	33-	99	44-	117	56-	138	69-	158	84-	176	99-	198	115-	221	133-	244	151-	269	171-	294	191-	321	213-	348	235-	377	259-	406	283-	437	309-	468	335-	501	363-	534	391-	569	421-	604
16	15-	69	24-	86	34-	104	46-	122	58-	142	72-	167	88-	184	102-	206	119-	229	138-	254	155-	279	175-	305	196-	332	218-	360	241-	389	265-	419	290-	450	315-	483	342-	516	370-	550	399-	585	429-	621
17	16-	72	25-	90	36-	108	47-	128	60-	148	74-	169	89-	191	105-	214	122-	238	140-	263	160-	288	180-	315	201-	343	223-	372	246-	402	271-	432	296-	464	322-	497	350-	530	378-	565	407-	601	437-	638
18	16-	76	26-	94	37-	113	49-	133	62-	154	76-	176	92-	198	108-	222	125-	247	144-	272	164-	298	184-	326	206-	354	228-	384	252-	414	277-	445	302-	478	329-	511	357-	545	385-	581	415-	617	446-	654
19	17-	79	27-	98	38-	118	50-	139	64-	160	78-	183	94-	206	111-	230	129-	255	148-	281	168-	308	189-	336	211-	365	234-	395	258-	426	283-	458	309-	491	336-	525	364-	560	393-	596	423-	633	454-	671
20	18-	83	28-	102	39-	123	52-	144	66-	166	81-	189	97-	213	114-	238	133-	264	152-	290	173-	318	193-	347	216-	376	239-	407	263-	439	289-	471	315-	505	343-	539	371-	575	401-	611	431-	649	463-	687
21	18-	86	29-	106	40-	128	53-	150	68-	172	83-	196	99-	221	117-	246	136-	272	155-	300	176-	328	198-	357	220-	388	244-	419	269-	451	295-	484	322-	518	350-	553	378-	590	408-	627	439-	665	471-	704
22	19-	89	29-	111	42-	132	55-	155	70-	178	85-	203	102-	228	120-	254	139-	281	159-	309	180-	338	202-	368	225-	399	250-	430	275-	463	301-	497	328-	532	356-	568	386-	604	416-	642	447-	681	480-	720
23	19-	93	30-	115	43-	137	57-	160	72-	185	87-	209	105-	235	123-	262	142-	290	163-	318	184-	348	207-	378	230-	410	255-	442	280-	476	307-	510	335-	545	363-	582	393-	619	424-	657	455-	697	488-	737
24	20-	96	31-	119	44-	142	58-	166	73-	191	90-	216	107-	243	126-	270	146-	298	167-	327	188-	358	211-	389	235-	421	260-	454	286-	488	313-	523	341-	559	370-	596	400-	634	431-	673	464-	712	497-	753
25	20-	100	33-	123	45-	147	60-	171	75-	197	92-	223	110-	250	129-	278	149-	307	170-	337	193-	367	216-	399	240-	432	265-	466	292-	500	319-	536	348-	572	377-	610	408-	648	439-	688	472-	728	505-	770

$\alpha = 0.02$

N_2	4		5		6		7		8		9		10		11		12		13		14		15		16		17		18		19		20		21		22		23		24		25	
N_1	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
4	—	—	15-	53	22-	64	29-	55	38-	66	48-	78	58-	92	70-	106	83-	121	96-	138	111-	155	127-	173	143-	193	161-	213	180-	234	199-	257	220-	280	242-	304	264-	330	288-	356	313-	383	338-	412
5	10-	30	16-	39	23-	49	31-	60	40-	72	50-	85	61-	99	73-	114	86-	130	100-	147	115-	165	131-	184	148-	204	166-	225	185-	247	205-	270	226-	294	248-	319	271-	345	295-	372	320-	400	345-	429
6	11-	33	17-	43	24-	54	32-	66	42-	78	52-	92	63-	107	75-	123	89-	139	103-	157	118-	176	135-	195	152-	216	171-	237	190-	260	210-	284	232-	308	254-	334	277-	362	302-	388	328-	417	354-	446
7	11-	37	18-	47	25-	59	34-	71	43-	85	54-	99	66-	114	78-	131	92-	148	107-	166	122-	186	139-	206	157-	227	176-	248	195-	271	216-	297	238-	322	261-	348	284-	376	308-	404	333-	433	359-	464
8	12-	40	19-	51	27-	63	35-	77	45-	91	56-	106	68-	122	81-	139	95-	157	111-	175	127-	195	144-	216	162-	238	181-	261	201-	285	222-	310	244-	336	267-	363	291-	391	316-	420	342-	451	370-	482
9	13-	43	20-	55	28-	68	37-	82	47-	97	59-	112	71-	129	84-	147	99-	165	116-	183	131-	203	148-	227	167-	249	188-	271	209-	297	228-	325	250-	350	274-	377	298-	406	324-	434	350-	466	378-	497
10	13-	47	21-	59	29-	73	39-	87	49-	103	61-	119	74-	136	88-	154	102-	174	118-	194	135-	213	153-	237	172-	260	191-	285	212-	310	234-	336	257-	361	282-	388	308-	416	333-	444	359-	482	388-	514
11	14-	50	22-	63	30-	78	40-	93	51-	109	63-	126	77-	143	91-	162	106-	182	122-	203	139-	223	157-	248	177-	271	197-	296	218-	322	240-	349	263-	376	288-	403	313-	433	339-	463	366-	494	394-	526
12	15-	53	23-	67	32-	82	42-	98	53-	115	66-	132	79-	151	94-	170	109-	191	126-	212	143-	233	162-	258	182-	282	202-	308	224-	334	246-	362	270-	390	295-	418	321-	447	347-	477	374-	508	403-	540
13	15-	57	24-	71	33-	87	44-	103	56-	120	68-	139	82-	158	97-	178	113-	199	130	221	148	244	167	268	187	293	208	319	230	346	253	374	276	404	303	434	329	464	356	495	383	526	411	564
14	16-	60	25-	75	34-	92	45-	109	58-	126	71-	145	85-	165	100-	186	116-	208	134-	230	152-	254	171-	279	192-	304	213-	331	235-	359	258-	387	283-	417	308-	448	335-	479	362-	511	391-	543	420-	576
15	17-	63	26-	79	36-	96	47-	114	60-	132	73-	152	88-	172	103-	194	120-	216	138-	238	156-	260	176-	289	197-	315	218-	343	241-	371	265-	400	290-	430	315-	462	342-	494	370-	527	398-	559	428-	592
16	17-	67	27-	83	37-	101	49-	119	62-	138	76-	158	91-	179	107-	201	124-	224	142-	248	161-	273	181-	299	202-	326	224-	354	247-	383	271-	413	296-	444	323-	475	350-	508	378-	541	407-	573	437-	608
17	18-	70	28-	87	39-	105	51-	124	64-	144	78-	165	93-	187	110-	209	127-	231	146-	257	165-	281	185-	310	207-	337	229-	366	253-	395	278-	425	303-	457	331-	488	359-	521	388-	555	417-	588	446-	624
18	19-	73	29-	91	40-	110	52-	130	66-	150	81-	171	96-	194	113-	217	131-	241	150-	266	169-	293	190-	320	212-	348	235-	377	259-	407	284-	438	310-	470	338-	503	367-	537	396-	572	426-	604	457-	640
19	19-	77	30-	95	41-	115	54-	135	68-	156	83-	178	99-	201	116-	225	134-	250	154-	275	174-	302	195-	330	217-	359	241-	388	265-	419	290-	451	317-	483	346-	516	375-	551	405-	586	435-	619	467-	653
20	20-	80	31-	99	43-	119	56-	140	70-	162	85-	185	102-	209	119-	233	138-	258	157-	285	178-	312	200-	340	222-	370	246-	400	271-	431	297-	463	323-	497	351-	531	380-	566	410-	602	441-	639	473-	677
21	21-	83	32-	103	44-	124	58-	145	72-	168	88-	191	105-	215	125-	240	142-	266	161-	294	182-	322	204-	351	228-	380	252-	411	277-	443	303-	476	330-	510	359-	544	388-	580	418-	617	449-	655	482-	693
22	21-	87	33-	107	45-	129	59-	151	74-	174	90-	198	108-	222	128-	248	145-	273	165-	303	187-	331	209-	361	233-	391	257-	423	283-	455	310-	488	337-	523	366-	558	395-	595	426-	632	458-	670	490-	710
23	22-	90	34-	111	47-	133	61-	156	76-	180	93-	204	110-	230	129-	256	149-	283	169-	312	191-	341	214-	371	238-	402	263-	434	289-	467	316-	501	344-	536	373-	572	403-	609	434-	647	466-	686	499-	726
24	23-	93	35-	115	48-	138	63-	161	78-	186	95-	211	113-	237	132-	264	152-	292	173-	321	196-	350	219-	381	243-	413	269-	445	295-	479	322-	514	351-	549	380-	586	411-	623	442-	662	475-	701	508-	742
25	23-	97	36-	119	50-	142	64-	167	81-	193	98-	217	116-	244	135-	272	156-	300	177-	330	200-	360	224-	391	248-	424	274-	457	301-	491	329-	526	358-	562	388-	599	418-	638	450-	677	483-	717	517-	758

(continued)

TABLE A-5 (continued)

$\alpha = 0.05$

N_1	4		5		6		7		8		9		10		11		12		13		14		15		16		17		18		19		20		21		22		23		24		25	
N_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
4	10-	36	16-	34	23-	43	31-	53	40-	64	49-	77	60-	90	72-	104	85-	119	99-	135	114-	152	130-	170	147-	189	164-	210	183-	231	203-	253	224-	276	246-	300	269-	325	293-	351	317-	378	343-	407
5	11-	29	17-	38	24-	48	33-	58	42-	70	52-	83	63-	97	75-	112	88-	127	103-	144	118-	162	136-	181	155-	201	175-	221	199-	243	209-	266	230-	290	253-	314	276-	340	300-	367	325-	395	352-	423
6	12-	32	18-	42	26-	52	34-	64	44-	76	55-	89	66-	104	79-	119	92-	136	107-	153	122-	172	139-	191	157-	211	175-	233	195-	253	215-	279	237-	301	260-	328	283-	355	308-	383	333-	411	360-	440
7	13-	35	20-	45	27-	57	36-	69	46-	82	57-	96	69-	111	82-	127	96-	144	111-	163	127-	181	144-	201	162-	222	181-	244	201-	267	222-	291	244-	314	267-	342	291-	369	316-	397	342-	426	369-	456
8	14-	38	21-	49	29-	61	38-	74	49-	87	60-	102	72-	118	85-	135	100-	152	115-	171	132-	191	149-	211	167-	233	187-	255	207-	279	230-	304	251-	329	274-	356	298-	384	324-	413	350-	442	378-	472
9	14-	42	22-	53	31-	65	40-	79	51-	93	62-	109	75-	125	89-	142	104-	160	119-	180	136-	200	154-	221	173-	243	192-	267	213-	291	237-	318	258-	342	281-	370	306-	398	332-	427	359-	457	387-	488
10	15-	45	23-	57	32-	70	42-	84	53-	99	65-	115	78-	132	92-	150	107-	169	124-	188	141-	209	159-	231	178-	254	198-	278	219-	303	242-	328	265-	355	289-	383	314-	412	340-	442	368-	472	396-	504
11	16-	48	24-	61	34-	74	44-	89	55-	105	68-	121	81-	139	96-	157	111-	177	128-	197	145-	219	164-	241	185-	265	204-	289	226-	314	248-	341	272-	368	296-	397	322-	426	349-	456	376-	486	405-	520
12	17-	51	26-	64	35-	79	46-	94	58-	110	71-	127	84-	146	99-	165	115-	185	132-	206	150-	228	169-	251	189-	275	210-	300	232-	326	255-	353	279-	381	304-	410	330-	440	357-	471	385-	503	414-	536
13	18-	54	27-	68	37-	83	48-	99	60-	116	73-	134	88-	152	103-	172	119-	193	136-	215	153-	237	174-	261	193-	285	216-	311	239-	337	262-	365	286-	394	312-	423	338-	454	365-	486	394-	518	423-	552
14	19-	57	28-	72	38-	88	50-	104	62-	122	76-	140	91-	159	106-	180	123-	201	141-	223	160-	246	179-	271	200-	296	222-	322	245-	349	269-	377	293-	407	319-	437	346-	468	374-	500	403-	533	433-	567
15	20-	60	29-	76	40-	92	52-	109	65-	127	79-	146	94-	166	110-	187	127-	209	145-	232	164-	256	184-	281	206-	306	228-	333	251-	361	275-	390	301-	419	327-	450	354-	482	382-	515	412-	548	442-	583
16	21-	63	30-	80	42-	96	54-	114	67-	133	82-	152	97-	173	113-	195	131-	217	150-	240	169-	265	190-	290	211-	317	234-	344	258-	372	282-	402	308-	432	335-	463	362-	496	391-	529	421-	563	451-	599
17	21-	67	32-	83	43-	101	56-	119	70-	138	84-	159	100-	180	117-	202	135-	225	154-	249	174-	274	195-	300	217-	327	240-	355	264-	384	289-	414	315-	445	342-	477	370-	510	399-	544	429-	579	461-	614
18	22-	70	33-	87	45-	105	58-	124	72-	144	87-	165	103-	187	121-	209	139-	233	159-	257	179-	283	200-	310	223-	337	246-	366	271-	395	296-	426	322-	458	350-	490	378-	524	408-	558	438-	594	470-	630
19	23-	73	34-	91	46-	110	60-	129	74-	150	90-	171	107-	193	124-	217	143-	241	163-	266	184-	292	205-	320	228-	348	252-	377	277-	407	303-	438	330-	470	358-	503	387-	537	416-	573	447-	609	479-	646
20	24-	76	35-	95	48-	114	62-	134	77-	155	93-	177	110-	200	128-	224	147-	249	167-	275	189-	302	211-	329	234-	358	258-	388	283-	417	310-	450	337-	483	365-	517	395-	551	425-	587	456-	624	489-	661
21	25-	79	37-	98	50-	118	64-	139	79-	161	95-	184	113-	207	131-	232	151-	257	172-	283	193-	311	216-	339	240-	368	264-	399	290-	430	317-	462	344-	496	373-	530	403-	565	434-	601	465-	639	498-	677
22	26-	82	38-	102	51-	123	66-	144	81-	167	98-	190	116-	214	135-	239	155-	265	176-	292	198-	320	221-	349	245-	379	270-	410	296-	442	326-	474	352-	508	381-	543	411-	579	442-	616	474-	654	508-	692
23	27-	85	39-	106	53-	127	68-	149	84-	172	101-	196	119-	221	139-	246	159-	273	180-	301	203-	329	226-	359	251-	389	276-	421	303-	453	330-	487	359-	521	389-	556	419-	593	451-	630	483-	669	517-	708
24	27-	89	40-	110	54-	132	70-	154	86-	178	104-	202	123-	227	142-	254	163-	281	185-	309	208-	338	232-	368	257-	399	282-	432	309-	465	337-	499	366-	534	396-	570	427-	607	459-	645	492-	684	527-	723
25	28-	92	42-	113	56-	136	72-	159	89-	183	107-	208	126-	234	146-	261	167-	289	189-	318	213-	347	237-	378	262-	410	289-	442	316-	476	346-	511	376-	546	404-	583	436-	620	468-	659	502-	698	536-	739

$\alpha = 0.10$

N_1	4		5		6		7		8		9		10		11		12		13		14		15		16		17		18		19		20		21		22		23		24		25	
N_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
4	11-	25	17-	33	24-	42	32-	52	41-	63	51-	75	62-	88	74-	102	87-	117	101-	133	116-	150	132-	169	150-	189	168-	208	187-	227	207-	249	228-	272	250-	296	273-	321	297-	347	322-	374	348-	402
5	12-	28	19-	36	26-	46	34-	57	44-	68	54-	81	66-	94	78-	109	91-	125	106-	141	121-	159	138-	177	155-	197	173-	218	193-	239	213-	262	235-	285	257-	310	281-	333	305-	362	330-	390	357-	418
6	13-	31	20-	40	28-	50	36-	62	46-	74	57-	87	69-	101	82-	116	95-	133	110-	150	126-	168	143-	187	161-	207	179-	229	199-	251	220-	274	242-	298	265-	323	289-	349	313-	377	339-	405	366-	434
7	14-	34	21-	44	29-	55	39-	66	49-	79	60-	93	72-	108	85-	124	99-	141	113-	159	131-	177	148-	197	166-	218	188-	239	206-	262	227-	286	249-	311	272-	337	297-	363	322-	391	348-	420	375-	450
8	15-	37	23-	47	31-	59	41-	71	51-	85	63-	99	75-	115	89-	131	104-	148	119-	167	136-	186	153-	207	172-	228	192-	250	212-	274	234-	298	257-	323	280-	350	305-	377	330-	406	357-	435	385-	465
9	16-	40	24-	51	33-	63	43-	76	54-	90	66-	105	79-	121	93-	138	108-	156	124-	174	141-	193	159-	214	178-	238	198-	261	219-	285	241-	310	264-	336	288-	363	313-	391	339-	420	366-	450	394-	481
10	17-	43	26-	54	35-	67	45-	81	56-	96	69-	111	82-	128	97-	145	112-	164	128-	184	146-	204	164-	226	184-	248	204-	272	226-	297	248-	322	272-	348	296-	376	321-	403	348-	434	375-	465	403-	497
11	18-	46	27-	58	37-	71	47-	86	59-	101	72-	117	86-	134	100-	153	116-	172	133-	192	151-	213	170-	235	190-	258	210-	283	232-	308	255-	334	279-	361	304-	389	329-	419	356-	449	384-	480	413-	512
12	19-	49	28-	62	38-	76	49-	91	62-	106	75-	123	89-	141	104-	160	120-	180	138-	200	156-	222	175-	245	196-	268	217-	293	239-	319	262-	346	286-	374	312-	402	338-	432	365-	463	393-	495	425-	527
13	20-	52	30-	65	40-	80	52-	95	64-	112	78-	129	92-	148	108-	167	125-	187	142-	209	161-	231	181-	254	201-	279	223-	304	245-	331	269-	358	294-	386	320-	415	348-	444	374-	477	403-	509	433-	542
14	21-	55	31-	69	42-	84	54-	100	67-	117	81-	133	96-	154	112-	174	129-	193	147-	217	166-	240	186-	264	207-	289	229-	315	252-	342	276-	370	301-	399	328-	428	355-	459	383-	491	412-	524	442-	558
15	22-	58	33-	72	44-	88	56-	105	69-	122	84-	141	99-	161	116-	181	133-	203	152-	225	171-	249	191-	274	213-	299	235-	326	259-	353	284-	381	309-	411	336-	441	363-	473	392-	505	422-	538	452-	573
16	24-	60	34-	76	46-	92	58-	110	72-	128	87-	147	103-	167	120-	188	138-	210	156-	234	176-	258	197-	283	219-	309	242-	336	266-	364	291-	393	317-	423	346-	454	372-	486	401-	519	431-	553	462-	588
17	25-	63	35-	80	47-	97	61-	114	75-	133	90-	153	106-	174	123-	195	142-	218	161-	242	181-	267	202-	293	225-	319	248-	347	273-	375	298-	405	325-	435	352-	467	380-	500	410-	533	440-	568	472-	603
18	26-	66	37-	83	49-	101	63-	119	77-	139	93-	159	110-	180	127-	203	146-	226	165-	251	186-	276	208-	302	231-	329	255-	357	280-	386	305-	417	332-	448	360-	480	389-	513	419-	547	450-	582	482-	618
19	27-	69	38-	87	51-	105	65-	124	80-	144	96-	165	113-	187	131-	210	150-	234	170-	259	191-	285	214-	311	237-	339	261-	368	286-	398	313-	428	340-	460	368-	493	398-	526	428-	561	459-	597	492-	633
20	28-	72	40-	90	53-	109	67-	129	83-	149	99-	171	117-	193	135-	217	154-	242	175-	267	196-	294	219-	321	243-	349	268-	378	293-	409	320-	440	348-	472	376-	506	406-	540	437-	575	469-	611	501-	649
21	29-	75	41-	94	55-	113	69-	134	85-	155	102-	177	120-	200	139-	224	159-	249	180-	275	202-	302	225-	330	249-	359	274-	389	300-	420	327-	452	355-	485	385-	518	415-	553	446-	589	478-	626	511-	664
22	30-	78	43-	97	57-	117	72-	138	88-	160	105-	183	123-	207	142-	232	163-	257	184-	284	207-	311	230-	340	255-	369	280-	400	307-	431	335-	463	363-	497	393-	531	423-	567	455-	603	488-	640	521-	679
23	31-	81	44-	101	58-	122	74-	143	90-	166	108-	189	127-	213	146-	239	167-	265	189-	292	212-	320	236-	349	261-	379	287-	410	314-	442	342-	475	371-	509	401-	544	432-	580	464-	617	497-	655	531-	694
24	32-	84	45-	105	60-	126	76-	148	93-	171	111-	195	130-	220	150-	246	171-	273	194-	300	217-	329	242-	358	267-	389	293-	421	321-	453	349-	487	379-	521	409-	557	441-	593	473-	631	507-	669	541-	709
25	33-	87	47-	108	62-	130	78-	153	96-	176	114-	201	133-	227	154-	253	176-	280	199-	308	222-	338	247-	368	273-	399	300-	431	328-	464	357-	498	386-	534	417-	570	449-	607	482-	645	516-	684	551-	724

*The data of this table are extracted with kind permission from DOCUMENTA GEIGY SCIENTIFIC TABLES, 6th Ed., pp. 124-127, Geigy Pharmaceuticals, Division of Geigy Chemical Corporation, Ardsley, N. Y.

APPENDIX B

FITTING DISTRIBUTIONS TO DATA

FITTING DISTRIBUTIONS TO DATA

Many of the outlier tests in Section 3 assume that the underlying distribution of the data is known. Four distributions which often provide good fits to air pollution and meteorological data are the normal, lognormal, exponential, and Weibull. The mathematical equations of these four distributions are:

$$\text{Normal:}\quad G(x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} \exp\,(-t^2/2)\,dt \qquad \text{B-1}$$

$$w = \frac{x - \mu}{\sigma} \qquad \text{B-2}$$

$$\text{Lognormal:}\quad G(x) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} \exp\,(-t^2/2)\,dt \qquad \text{B-3}$$

$$w = \frac{\ln x - \mu}{\sigma} \qquad \text{B-4}$$

$$\text{Exponential:}\quad G(x) = \exp\,[-\lambda(x-\theta)] \qquad \text{B-5}$$

$$\text{Weibull:}\quad G(x) = \exp\,[-(x/\delta)^{k}]. \qquad \text{B-6}$$

G(x) is defined as the fraction of a population having values greater than x.

Each of the four distributions described above can be completely described by specifying values for two parameters. These parameters relate to the shape, scale, or location of the distribution when plotted on graph paper. Accurate estimates of the maximum values of a data set require good estimates of the parameters of the distribution which most closely fit the data. Two methods of fitting these distributions are discussed in this section: the traditional method of maximum likelihood which uses all the data and a least squares approach, which emphasizes the upper tail of the data.

THE METHOD OF MAXIMUM LIKELIHOOD

The principal advantage of using the method of maximum likelihood to estimate distribution parameters is that it produces estimates which have minimum variance and distributions which asymptotically approach the normal distribution. Table B-1 lists maximum likelihood estimators (MLE's) for the parameters of the four distributions discussed above. Unfortunately, the MLE's for the shape and scale parameters of the Weibull distribution cannot be calculated directly. The equations

$$\hat{k} = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{\delta}}\right]^{\frac{1}{\hat{\delta}}} \qquad \text{B-7}$$

$$\hat{\delta} = n\left[\left(\frac{1}{\hat{k}}\right)^{\hat{\delta}}\sum_{i=1}^{n} x_i^{\hat{\delta}} \ln x_i - \sum_{i=1}^{n} \ln x_i\right]^{-1} \qquad \text{B-8}$$

must be solved simultaneously using an iterative procedure.[1] A computer program which uses a "golden section" iterative procedure has been developed by Mage et al.[2]

THE METHOD OF LEAST SQUARES

The parameters estimated by the maximum likelihood method determine distributions which fit the whole data set. However, the maximum value of a given data set is often better estimated using the parameters of a distribution fit to the upper tail of the data distribution by the method of least squares. This method requires that the equation defining the distribution under consideration be expressed as a linear relationship of the form $z = ay + b$. Equations B-1, B-3, B-5, and B-6 can be rewritten in linear form if the following identities are used.

Distribution	z	a	y	b
normal	z	$\frac{1}{\sigma}$	x	$-\frac{\mu}{\sigma}$
lognormal	z	$\frac{1}{\sigma}$	ln x	$-\frac{\mu}{\sigma}$
exponential	ln G(x)	$-\lambda$	x	$\lambda\theta$
Weibull	ln[-ln G(x)]	k	ln x	$-k\ln\delta$

The values of z for the normal and lognormal distributions are determined from the standard normal distribution such that the area under the standard normal curve from z to ∞ is equal to G(x). The following table lists z values for selected G(x) values in the upper tail of the data distribution.

G(x)	z	G(x)	z
0.50	0	0.02	2.054
0.40	0.253	0.01	2.326
0.30	0.524	0.005	2.576
0.20	0.842	0.002	2.880
0.10	1.282	0.001	3.090
0.05	1.645	0.0005	3.291

A linear regression analysis of the data which have been transformed from x and G(x) values to y and z by the appropriate identities listed above will determine a regression line which has an equation in the form of $z = ay + b$. Parameters of the corresponding distribution can be determined from the values of a and b using the equations listed in Table B-1 under least squares estimators (LSE's).

One measure of the goodness of fit of the distribution under consideration is the coefficient of determination (r^2), which is determined as part of the linear regression analysis. The closer the r^2 value is to unity, the better the distribution fits the data.

There are numerous other statistics which have been suggested for quantifying goodness of fit. EPA has developed a program which calculates six goodness of fit statistics: absolute deviations, weighted absolute deviations, Chi-square, Kolmogov-Smirnov, Cramer-von Mises-Smirnov, and the maximum value of the log-likelihood function.[2] Four other goodness of fit statistics in common use are the Kuiper, Watson, Anderson-Darling, and Shapiro-Wilk statistics.[3] Of these statistics, the r^2 and Chi-square are the easiest to calculate, though not necessarily the best choices for distinguishing which of the distributions under investigation provide the best fit to the data.

However, for routine data validation procedures requiring the selection of a distribution to characterize the data, the r^2 statistic is recommended since it can be easily calculated as part of the linear regression procedure. In general, the distribution which yields the highest r^2 value should be used to characterize the data. Parameter values are then estimated using the LSE equations in Table B-1.

TABLE B-1. ESTIMATION OF DISTRIBUTION PARAMETERS

Distribution	Parameter			Estimator	
	Location	Scale	Shape	MLE	LSE
Normal	μ	σ	--	$\hat{\mu} = \bar{x}$	$\hat{\mu} = \frac{-b}{a}$
				$\hat{\sigma} = s$	$\hat{\sigma} = \frac{1}{a}$
Lognormal	--	μ	σ	$\hat{\mu}$ = mean of ln x	$\hat{\mu} = \frac{-b}{a}$
				$\hat{\sigma}$ = std. dev. of ln x	$\hat{\sigma} = \frac{1}{a}$
Exponential	θ	λ	--	$\hat{\theta} = \min x_i$	$\hat{\theta} = \frac{-b}{a}$
				$\hat{\lambda} = \frac{n}{\Sigma(x_i - \min x_i)}$	$\hat{\lambda} = -a$
Weibull	--	δ	k	Iterative solutions of simultaneous equations B-7 and B-8	$\delta = \exp(-b/a)$ $k = a$

REFERENCES

1. Hastings, N. A. J., and J. B. Peacock. Statistical Distributions. John Wiley and Sons, New York.

2. Mage, David T., et al. "Techniques for Fitting Probability Models to Experimental Data," In: Proc. of Specialty Conference on Quality Assurance in Air Pollution Measurement, New Orleans, Louisiana, March 1979.

3. Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. Journal of the American Statistical Association, Vol. 69, No. 347, September 1974.

APPENDIX C

CALCULATION OF LIMITS FOR SHEWHART CONTROL CHART

CALCULATION OF LIMITS FOR SHEWHART CONTROL CHART

The Shewhart control chart enables the data analyst to test the hypothesis H_o that the mean or range value of the sample under evaluation comes from a population having the same distribution as the historical data sets A_1, A_2, ..., A_n.

The first step of the procedure is the selection of a suitable sample size for the test, as indicated in Section 3.3.3. Each sample should contain between 4 and 15 values and should represent a well-defined time period (day, month, quarter, etc.) for which there is a large body of historical data. Where possible these time periods should relate to the NAAQS of interest. Months or quarters would be appropriate time periods for tests of 24-hour TSP, SO_2, and NO_2 data collected at 6- or 12-day intervals.

The second step is the selection of historical data for determining the limits on the Shewhart control chart. To the extent possible, these data sets should contain data collected under the same conditions (averaging time, site location, season, weather, local emissions, etc.) as the data set under investigation. For convenience, these data sets will be labeled A_1, A_2, ..., A_N. Data set A_1 contains n_1 values and has a mean $\bar{x}_1$ and a range R_1. Similarly, data set A_2 contains n_2 values and has a mean $\bar{x}_2$ and a range R_2. Continuing in this manner, the data analyst can develop the following table.

Data set	Sample size	Mean	Range	d_2	$\frac{\text{Range}}{d_2}$
A_1	n_1	$\bar{x}_1$	R_1	d_{21}	R_1/d_{21}
A_2	n2	$\bar{x}_2$	R_2	d_{22}	R_2/d_{22}
⋮					
A_N	n_N	$\bar{X}_N$	R_N	d_{2N}	R_N/d_{2N}

TABLE C-1. FACTORS FOR ESTIMATING CONTROL LIMITS OF SHEWHART CHART

Number of observations in subgroup n	Factors[a]		
	d_2	c_2	c_2/d_2
2	1.128	0.5642	0.5002
3	1.693	0.7236	0.4274
4	2.059	0.7979	0.3875
5	2.326	0.8407	0.3614
6	2.534	0.8686	0.3428
7	2.704	0.8882	0.3285
8	2.847	0.9027	0.3171
9	2.970	0.9139	0.3077
10	3.078	0.9227	0.2998
11	3.173	0.9300	0.2931
12	3.258	0.9359	0.2873
13	3.336	0.9410	0.2821
14	3.407	0.9453	0.2775
15	3.472	0.9490	0.2733
16	3.532	0.9523	0.2696
17	3.588	0.9551	0.2662
18	3.640	0.9576	0.2631
19	3.689	0.9599	0.2602
20	3.735	0.9619	0.2575
21	3.778	0.9638	0.2551
22	3.819	0.9655	0.2528
23	3.858	0.9670	0.2506
24	3.895	0.9684	0.2486
25	3.931	0.9696	0.2467
30	4.086	0.9748	0.2386
35	4.213	0.9784	0.2322
40	4.322	0.9811	0.2270
45	4.415	0.9832	0.2227
50	4.498	0.9849	0.2190
55	4.572	0.9863	0.2157
60	4.639	0.9874	0.2128
65	4.699	0.9884	0.2103
70	4.755	0.9892	0.2080
75	4.806	0.9900	0.2060
80	4.854	0.9906	0.2041
85	4.898	0.9912	0.2024
90	4.939	0.9916	0.2008
95	4.978	0.9921	0.1993
100	5.015	0.9925	0.1979

[a]These factors assume sampling from a normal universe.

Figure 3-7 of Section 3 shows the general form of a control chart for testing sample means. Values of $\bar{x}$ are indicated on the vertical axis and units of time (or sample number) are indicated on the horizontal axis. The central line is a solid horizontal line drawn at

$$\bar{\bar{x}} = \frac{1}{N} \sum_{i=1}^{N} \bar{x}_i . \qquad \text{C-1}$$

The upper and lower control limits are dashed horizontal lines drawn at

$$UCL_{\bar{x}} = \bar{\bar{x}} + z\sigma_{\bar{x}} \qquad \text{C-2}$$

and

$$LCL_{\bar{x}} = \bar{\bar{x}} - z\sigma_{\bar{x}} \qquad \text{C-3}$$

where

$$\sigma_{\bar{x}} \cong \frac{1}{N\sqrt{n_*}} \sum_{i=1}^{N} \frac{R_i}{d_{2i}} \qquad \text{C-4}$$

and n_* is the size of the sample being tested. Computations are considerably reduced if samples are selected such that $n_* = n_1 = n_2 = \ldots = n_N = n$. In this case, the second, fifth, and sixth columns in the table above can be omitted and

$$\sigma_{\bar{x}} \cong \frac{1}{d_2 N\sqrt{n}} \sum_{i=1}^{N} R_i . \qquad \text{C-5}$$

Figure 3-8 of Section 3 shows the general form of a control chart for testing sample ranges. Values of R are indicated on the vertical axis, and units of time (or sample number) are indicated on the horizontal axis. The central line is a solid horizontal line drawn at

$$\bar{R} = \frac{1}{N} \sum_{i-1}^{N} R_i \qquad \text{C-6}$$

The upper and lower control limits are dashed horizontal lines drawn at

$$UCL_R = \bar{R} + z\sigma_R \qquad \text{C-7}$$

$$LCL_R = \bar{R} - z\sigma_R \qquad \text{C-8}$$

where

$$\sigma_R \approx \frac{c_2}{N} \sum_{i=1}^{N} \frac{R_i}{d_{2i}} \qquad \text{C-9}$$

and c_2 is determined from Table C-1 according to the size n_* of the sample being tested. Negative LCL's should be treated as zeros. If samples are selected such that $n_* = n_1 = n_2 = \ldots = n_n$, then

$$\sigma_R \approx \frac{c_2}{d_2 N} \sum_{i=1}^{N} R_i = \frac{c_2 \bar{R}}{d_2} \qquad \text{C-10}$$

Values of $\frac{c_2}{d_2}$ for $0 < n < 100$ are listed in Table C-1.

The data analyst is testing the hypothesis H_o that the mean or range value of the sample under evaluation comes from a population having the same distribution as the historical data sets $A_1, A_2, \ldots, A_n$. If H_o is true, the probability of the mean or range value falling outside the control limits is assumed to be equal to twice the area under the standard normal curve to the right of z. Consequently, the z value contained in Equations C-2, C-3, C-7, and C-8 is selected according to the desired sensitivity of the test. The following is a list of probabilities corresponding to selected z values.

z	P[H_o is true]
1.282	0.20
1.645	0.10
1.96	0.05
2.00	0.0455
2.326	0.02
2.576	0.01
3.00	0.0027

The value z = 3 corresponding to P[H_o is true] $\leq$ 0.0027 is commonly used for materials testing, but it may be too large for testing air quality data. The data analyst may wish to use z = 2 to determine the initial control limits and to later increase z if the original limits flag too many valid data sets.